



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

**A computational exploration of a possible
alternative to nucleotides as the basis of a genetic
alphabet**

A thesis presented to the University of Dublin, Trinity College for
the degree of Doctor of Philosophy

2012

NíChaoimh Lavinia Dewdney



Thesis 9671

—

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university.

It is entirely my own work, except where otherwise cited, referenced, acknowledged or accredited.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Thesis Summary

One of the most fundamental questions in molecular biology is why nature has chosen Adenine (A), Cytosine (C), Guanine (G), Uracil (U)/Thymine (T) for the genetic alphabet. Although much is known about the structure and composition of DNA the reason behind nature's particular choice of nucleotide alphabet over the many conceivable alternatives is not self-evident. Most studies have pursued physicochemical aspects of the problem while informatics aspects have been largely neglected, although they have been recently shown to play a fundamental role.

In the familiar terrestrial genetic alphabet information is encoded both in the purine/pyrimidine nature of a nucleotide (1-bit), and in the hydrogen/lone-pair donor/acceptor (D/A) patterns expressible using up to three positions (3-bits). Inspired by the reverse-engineering approach of Eschenmoser, the potential viability of a molecular alphabet built from molecules other than nucleotides, but possessing the same inherent capacity to express information, is explored. Specifically, alphabets in which information is expressed using 4-bit donor-acceptor (D/A) patterns as opposed to 3-bit D/A patterns, plus 1-bit corresponding to size, are considered. By comparing and contrasting the properties of the familiar and alternative alphabets it is hoped to learn something about the engineering factors underlying the evolution of the familiar, terrestrial alphabet.

A potential 4-bit molecular alphabet based on a naphthalene skeleton was proposed with the various D/A patterns created by the introduction of heteroatoms, carbonyl groups and amino-groups as required; the informal term heteronaphthalene (contracted to Het) proved convenient in referencing this 'ideal' set. A second 4-bit set, designed to act as a control, was based as far as possible on molecules actually realised in the laboratory of Zimmerman et al. (termed Zim).

The primary goal of this study was to determine how molecules possessing the various 4-bit D/A patterns interacted in the absence of any distinguishing size feature. The chemical realisation of D/A would have been complicated by matters such as chemical and tautomeric stability, as well as synthetic accessibility, and these in turn would often depend on the particular chemical expression of the D/A patterns rather than on the patterns

themselves. Accordingly, a computational approach is adopted. For most purposes the well-known programs Gaussian 03W and Spartan 04 V1.0.1 are used to calculate the necessary molecular data. A variety of computational approximations were selected, ranging from the semi-empirical (AM1, PM3) to *ab initio* Hartree-Fock with MP2, so as to minimise the possibility of artefacts arising from the characteristics of any one method. Alphabet properties which are essentially common across the various computational approximations are considered likely to be reliable.

The role of D/A patterns in determining the viability of potential alphabets is twofold, serving to bind associating complementary pairs, while simultaneously opposing non-complementary associations. Any set of complementary D/A is approximately equivalent with respect to the former; however, molecular alphabets possessing greater resistance to non-complementary associations, thereby preserving information integrity and avoiding genetic error, should possess evolutionary advantage. The capacity to avoid errors is based on interaction energies as determined using the various computational methods. In short, it is observed that a subset of 6 of the 16 available letters {D, D*, F, F*, G, and G*} had strong mutual repulsions between non-complements, yielding a small potentially viable subset 6 letters. By contrast in nucleotides, and based on consideration of patterns alone, 8 of the 16 available patterns appear to be viable. In the Het alphabet the maximum mutual repulsion in the alphabet is in 2 of 4 D/A positions whereas in nucleotides it is in 2 of 3, and this difference can be directly related to the size feature in nucleotide but absent in the Het alphabet. The results for the Zim alphabet were similar although the alphabet was slightly reduced in size. This 'Eschenmoser' result informs us that pyrimidine/purine size asymmetry offers evolutionary advantage and is unlikely to be an accident of biochemistry. This is a significant result and confirms the theoretical prediction from the recently further developed error-coding model by Mac Dónaill.

Less significant, yet notable results include an analysis of the role of secondary interaction between adjacent H-bonds which successfully extends the work of Jorgensen et al. Possible effects of molecular flexibility are also briefly considered; three further potential alphabet sets are introduced. Nitrogen 'pyramidalization' is also examined and the results indicate that this may further restrict the size of the viable alphabets.

Acknowledgements

To my father

Stephen Maurice John Dewdney

I would like to thank my supervisor Professor Dónall Mac Dónaill, from whom I have learned so much, for all his help and for always pushing me to achieve more.

Much appreciation is due to all of my colleagues/friends in the computational research lab and throughout Trinity College Dublin. I am very grateful to all the staff in the School of Chemistry and the Trinity Centre for High Performance Computing (TCHPC). I am especially grateful to Professor Graeme Watson who always included me in the activities (academic and social) of his research group.

This thesis would not have been possible without the support of my mother Cecilia who encouraged me every step of the way and in times of self doubt pushed me forward. I owe so much thanks to my Fiancé Daithí for more than I could possibly mention, everything from proof reading to supplying me with chocolate to keep me going but most importantly for always reminding me I could reach the end goal.

The work in this thesis was funded by the HEA under a PRTLII (Cycle III) grant, a Trinity College Dublin postgraduate award and by funds from the School of Chemistry (for which I am very thankful to Professor John Corish). All calculations were performed on the IITAC cluster maintained by TCHPC.

Contents

1 Introduction	1
1.1 Understanding the building blocks of life	1
1.1.1 The structure of DNA	5
1.1.2 A reverse engineering approach	8
1.2 Error-coding theory	9
1.2.1 Parity	11
1.3 Nucleotide Donor/Acceptor Patterns	12
1.4 The advantage of code partitioning	15
1.5 Designing an alternative to nucleotides	19
2 Computational Theory and Methods	27
2.1 Introduction	27
2.1.2 Schrödinger Equation	27
2.1.3 Variation Theorem	28
2.1.4 Born Oppenheimer Approximation	28
2.2 Hartree-Fock Method	29
2.3 Basis Set Choice	31
2.4 Basis Set Superposition Error (BSSE)	32
2.5 Semi-empirical Methods	34
2.5.1 AM1 (Austin Model 1) and PM3 (Modified Neglect of Diatomic Overlap, Parametric Method 3)	36
2.6 Møller-Plesset Model	36
3 The Heteronaphthalene Potential Alphabet Letter Set	41
3.1 Designing a Heteronaphthalene Potential Alphabet	41
3.2 Exploring a potential Heteronaphthalene alphabet	45
3.2.1 Construction of molecular geometry constraints	47
3.3 Complementary Heteronaphthalene associations – Results	49

4 Secondary Interactions	53
4.1 Introduction	53
4.2 Heteronaphthalene secondary interactions	56
4.2.1 Heteronaphthalene based fit for secondary interactions	59
5 Non-complementary Heteronaphthalene Associations	67
5.1 Introduction	67
5.1.2 Imposing standard geometric constraints in mismatching positions	68
5.2 Non-complementary Heteronaphthalene associations-Results	69
5.2.1 Mismatches in one position	69
5.2.2 Mismatches in two positions	73
5.2.3 Mismatches in three positions	77
5.2.4 Mismatches in four positions	80
5.3 Heteronaphthalene HF results: Discussion and Conclusions	81
6 BSSE and Different Computational Methods	89
6.1 Introduction	89
6.2 Basis Set Superposition Error	89
6.3 BSSE calculation results	89
6.3.1 BSSE complementary associations	89
6.3.2 BSSE mismatches in one and three positions	90
6.3.3 BSSE mismatch in two positions	94
6.3.4 BSSE mismatch in four positions	96
6.4 Discussion and Summary-BSSE	98
6.5 Semi-empirical methods	100
6.5.1 Semi-empirical methods-Introduction	100
6.6 Semi-empirical Results	100
6.6.1 Complementary associations	100
6.6.2 Mismatches in one and three positions	101
6.6.3 Mismatches in two positions	108
6.6.4 Mismatches in four positions	113
6.7 Discussion and Summary-Semi-empirical results	114
6.8 MP2	117
6.8.1 Introduction	117

6.8.2 Non- Complementary Pairs	119
6.9 Summary and Discussion – MP2	125
7 The Zimmerman Potential Alphabet	127
7.1 Introduction	127
7.2 Zimmerman Results	130
7.2.1 Free complements associations	130
7.2.2 Molecular geometry constraints	133
7.3 Non-complementary associations	136
7.3.1 Mismatches in one and three positions	136
7.3.2 Mismatches in two positions	144
7.3.3 Mismatches in four positions	150
7.4 Summary and Discussion-Zimmerman results	152
8 The effects of geometric freedom and flexibility	157
8.1 Introduction	157
8.2 Nitrogen pyramidalization	158
8.2.1 Monomer pyramidalization	159
8.3 Pyramidalization-Results	160
8.3.1 Pyramidalized Heteronaphthalene monomers	160
8.3.2 Mismatches in one position	163
8.3.3 Mismatches in two positions	166
8.4 Summary and conclusions- Pyramidalized Heteronaphthalenes	166
8.5 Molecular flexibility	167
8.5.1 Adapted N-H---N Heteronaphthalenes [Hetnhn]	167
8.5.2 N-H---N Skeletal [Skelnhn]	170
8.5.3 Mixed Skeletal [Skelmix]	173
8.6 Discussion and conclusions-All data sets	178
9 Conclusions	183
9.1 Introduction	183
9.2 A Heteronaphthalene alphabet	183
9.3 A Zimmerman alphabet	186
9.4 Pyramidalization	188

9.5 Molecular flexibility	189
9.6 Future work	190
Appendices	193
Bibliography	215

1 Introduction

1.1 Understanding the building blocks of life

In 1953 Watson and Crick made an important breakthrough in determining the structure of Deoxyribonucleic Acid (DNA) [1], putting forward a double helix structure consisting of two chains coiled about the same axis. The two chains in the helical structure are linked together through hydrogen bonds between nucleotide base pairs A:T (Adenine (A) Thymine (T)) and C:G (Cytosine (C) Guanine (G)) (Fig. 1.1).

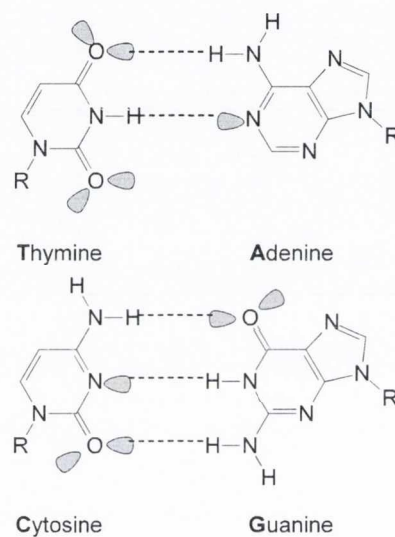


Fig. 1.1 Base pairs in DNA

Although the molecular composition and construction of DNA are now known important questions still remain regarding why DNA is as it is. Why are only four nucleotides used? A:T, C:G are not the only nucleotide pairs that exist, alternatives have been proposed (Fig. 1.2) and successfully incorporated into DNA (Switzer et al [2], Piccirilli et al [3]) proving that others pairs are at least possible. Since an extended nucleotide alphabet has been shown to be possible why does nature use only the four bases?

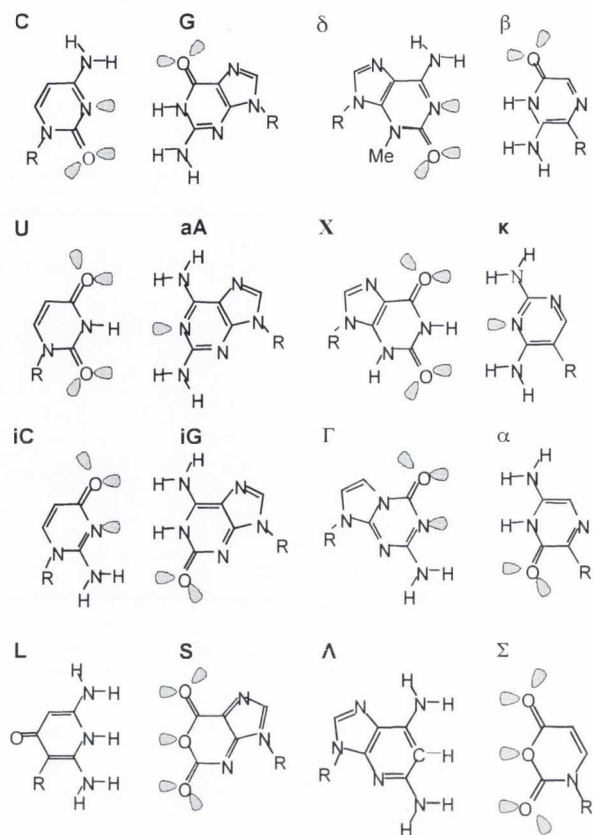


Figure 1.2 Extended nucleotide alphabet [3, 4].

Indeed, given the wide variety of molecules potentially capable of molecular recognition (some examples can be seen in Fig. 1.3a-c), the question arises as to why are nucleotides used as opposed to the many alternatives? Nature's particular choice of nucleotides is not self-evident.

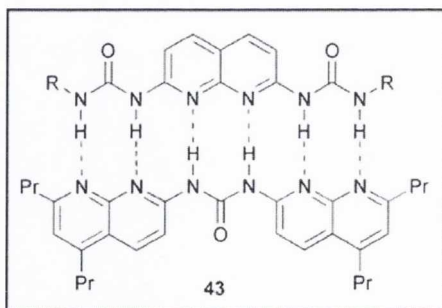


Figure 1.3a Structure with multiple hydrogen bonds.
Figure copied directly from reference Sijbesma[5].

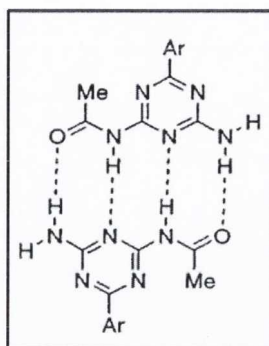


Figure 1.3b Structure with multiple hydrogen bonds Zimmerman and Corbin [6]

Figure copied directly from reference.

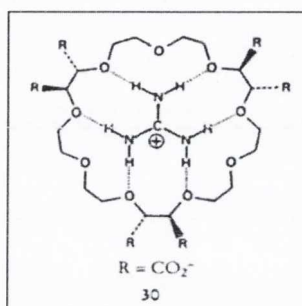


Figure 1.3c Structure with multiple hydrogen bond Lehn[7]

Figure copied directly from reference.

One of the many possibilities is that nature's particular choice is the result of a frozen accident and that if the prebiotic environment on early Earth had been different a different outcome could have occurred. Another distinct possibility is that the combination of A, T, C and G is superior to all others available meaning that even if the environment on early Earth had been different the most probable outcome would still be A, T, C and G. Albert Einstein suggested the general purpose;

"We not only want to know how nature is (and how her transactions are carried through), but we also want to reach, if possible, a goal which may seem utopian and presumptuous, namely to know why nature is such and not otherwise" [8].

Most explorations into the origins of life and into DNA more specifically have employed a physicochemical perspective (see references [9-16] for examples). Relatively few studies have considered DNA and its replication from the perspective of information transmission.

The link between information and life was put succinctly by Dawkins who said:

“If you want to understand life, don’t think about vibrant, throbbing gels and oozes, think about information technology” [17]

If DNA and its replication were to be considered not exclusively from a physicochemical perspective but in term of information and its transmission (governed by the rules that pertain to information) this could provide a possible insight into nature’s choice. Yockey suggested a potential role for information in nucleotides, assigning each a 5-bit numerical representation [18]. In his work Szathmáry reflected on the importance of hydrogen donor-acceptor (D/A) patterns and the role they play in replication fidelity [19] but did not combine this with the rules of information. Independently of Yockey and Szatháry these two perspectives were combined and taken further in the work of Mac Dónaill [20][21], who translated each nucleotide (the complete set of 16 possible nucleotides shown in Fig. 1.2) into a binary pattern based on the hydrogen donor/acceptor (D/A) (3-bits) pattern and the size (1-bit) of each molecule (purine or pyrimidine) (Fig. 1.4) (see section 1.3).

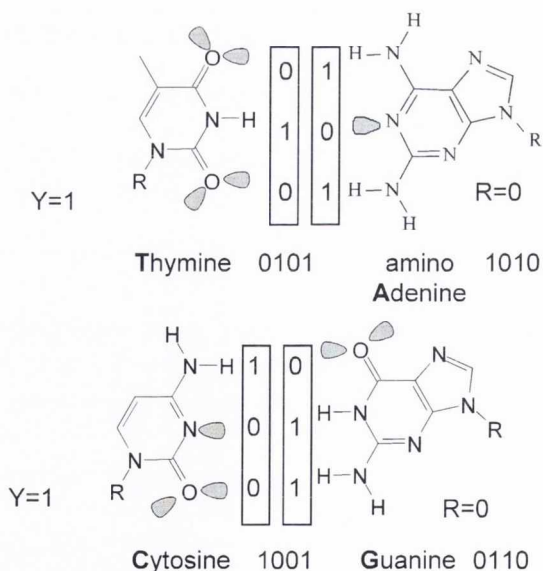


Figure. 1.4 Binary representation for base pairs TaA and CG.

D/A pattern, Hydrogen = 1 and Lone Pair (LP) = 0, 3-bits in total; size (4th-bit) , puRine R=0 and pYrimidine Y=1, 1bit

Through interpreting nucleotides in terms of digital patterns as opposed to purely chemical composition, Mac Dónaill has successfully shown with the use of error-coding theory that just as some combinations of patterns are superior to others so are some combinations of nucleotides. Mac Dónaill has extended his model and shown that alphabets containing only one molecular size are inherently inferior to those in which the molecules can be divided into groups based on size (see section 1.4).

This thesis will consider the possibility of a 4-bit D/A alphabet composed from molecules other than nucleotides. In doing this a reverse engineering approach will be adopted (section 1.1.2). This approach will consist of proposing, developing and modelling a possible alternative alphabet letter set composed of molecules other than nucleotides. In the investigations carried out in this thesis a computational/theoretical approach will be taken. By exploring an alternative in this way perhaps we can learn about what is as opposed to what could be.

Before considering the information carried in DNA or proposing an alternative to the nucleotide alphabet it is important to be familiar with the structure and key components of DNA and with the approach being undertaken in this work.

1.1.1 The structure of DNA

The reader will find a useful exploration of the theory presented here in [22,23]. Watson and Crick [1] proposed that the structure of DNA consists of two helical chains of polynucleotides, each coiled around the same axis. The four bases A, C, G, T in DNA form two pairs held together by hydrogen bonds. C is paired with G (held by three hydrogen bonds) and A is paired with T (held by two hydrogen bonds). In each base pair one molecule is a purine (A, G) and the other a pyrimidine (T, C) (Fig. 1.1). DNA is a polymer built from nucleotide repeating units. Each unit consists of the sugar deoxyribose, a base (the base and sugar unit is called a nucleoside) and a phosphate (nucleoside plus the phosphate group gives a nucleotide). Polymerisation occurs through the condensation of a phosphate group on one nucleotide unit with the hydroxyl group of the sugar on another nucleotide, thus, the separate nucleotide units are joined to each other through a phosphodiester bond (Fig. 1.5). As seen in the figure below (Fig. 1.5), a polynucleotide

has two distinct ends, 3' and 5'. At the 3' end a hydroxyl group is attached at the third carbon position of the sugar, whilst at the 5' end a phosphate group is attached to the 5' carbon position (for enlarged picture of the sugar see Fig. 1.6).

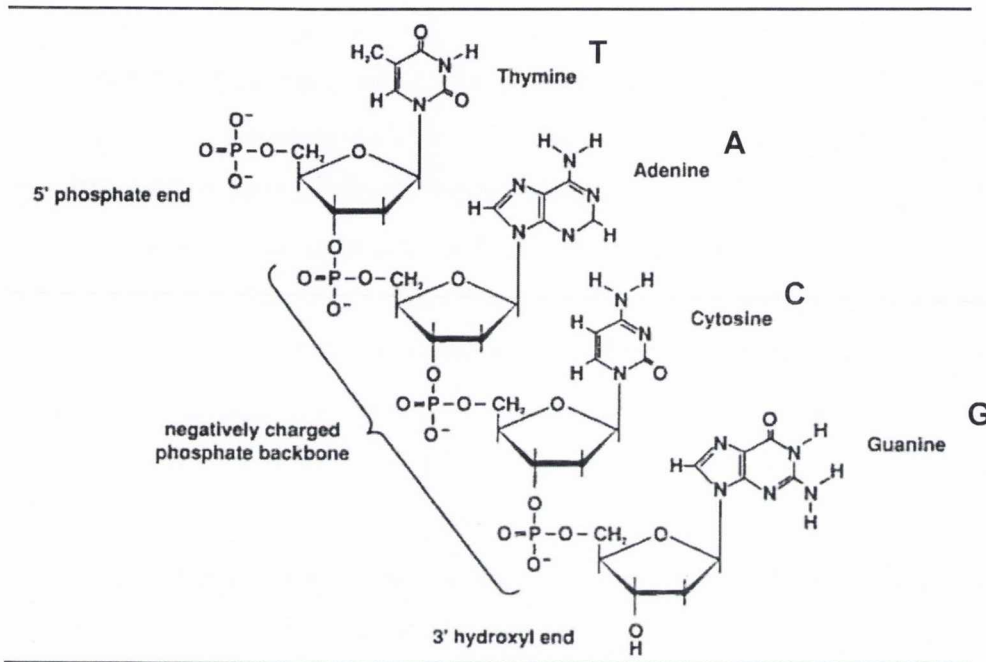


Figure 1.5 Single strand of DNA, showing polarity 5'→3' direction. Diagram adapted from [24]

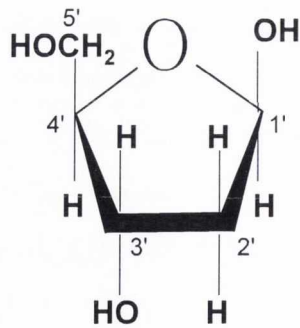


Figure 1.6 Deoxyribose with labelled carbon positions.

Each strand has a polarity as indicated in Fig. 1.6 (Note: convention dictates that strands are written starting at the 5' end). Two complementary strands of nucleotides are joined together by hydrogen bonding between the base pairs. The complementary strands are aligned antiparallel to each other. One strand will be 5'→3' direction and the other 3'→5'.

Double stranded DNA twists to form the familiar double helix structure. A double stranded helix can exist in different structural forms. B- DNA[25] for example, as discovered by Watson and Crick is right-handed (Fig. 1.7) but DNA can also be A form (right-handed)[25] or Z form (left-handed)[25]. Several parameters are used to describe the double helix structure such as the major and minor groove and pitch and rise (detailed in Fig. 1.7).

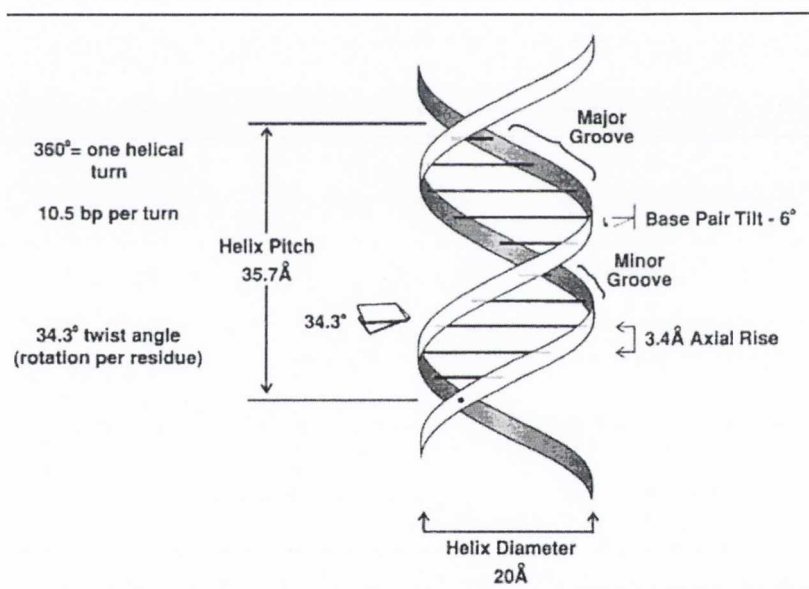


Figure 1.7 B- DNA double helix showing major and minor grooves, tilt, rise, twist and pitch. Diagram taken directly from [26]

Replication in DNA occurs with the aid of polymerase. The parent helix is unravelled and each pre-existing strand is used as a template for a strand in a daughter helix. In each new strand that is formed A will be replaced by T and C by G and visa versa. Polymerase allows growth only in the 5'→3' chain direction and propagates into the parent helix (this process can be likened to unzipping a closed zipper). For the strand in the parent helix with 5'→3' polarity (the leading strand) directionality of growth allowed by polymerase is not a problem and a daughter strand is readily formed. The remaining strand of the parent helix

with 3'→5' polarity (the lagging strand) must be replicated in short pieces known as Okazaki pieces in order to produce a strand with the correct polarity.

Although the mechanism and function of DNA is now well understood, the fundamental question as to why nucleotides are employed as opposed to potential alternatives remains little considered.

1.1.2 A reverse engineering approach

The reverse engineering approach adopted in this thesis is inspired by that seen in the work of Eschenmoser and summarized;

“The strategy is to conceive (through chemical reasoning) potentially natural alternatives to the nucleic acid structure, to synthesize such alternatives by chemical methods, and to compare them with the natural nucleic acids with respect to those chemical properties that are fundamental to the biological function of RNA and DNA.” [27]

He investigated why nature chose to use a specific pentose sugar in DNA (2-deoxyribose) by exploring a hexose alternative and studying the outcome [28, 29]. In order to learn about why the 2-deoxyribose is used in nature, Eschenmoser explored a plausible alternative, 2,3-dideoxy-glucofuranose [28] and observed the outcome. It was seen that changing the deoxyribose to a dideoxyhexopyranose caused deviation from classic Watson-Crick pairing and amongst other things made self interactions between A:A and G:G and increased base pairing strength [27]. Eschenmoser and Dobbler concluded that the pentose ring is primarily responsible for the helical structure of double stranded DNA [29].

DNA replication is responsible for transferring information from one generation of cells to the next; viewing nucleotide replication as an information process could potentially yield important information when searching for the answers behind nature's choice. Recognising that approaching questions relating to nature's choice of nucleotides from the point of information has been somewhat neglected and is unfamiliar to most we now consider some of the fundamental concepts.

1.2 Error-coding theory

Codes are frequently used in transmitting information from one point to another. A code can be defined as a set of codewords. The required information is first encoded, then transferred, and finally decoded at its destination. For this process to work the transmitter and receiver must share the same dictionary of codewords. If a valid dictionary word from the dictionary is received the receiver presumes this is error free and proceeds with the command. Coding gives the possibility that if a codeword (an element of a code) has been distorted during transmission, the distorted codeword could be detected and resent or, in some cases the codeword could even be corrected.

Two key concepts in determining the strength of a code are (i) the Hamming distance [30] between two codewords $\partial(a,b)$, which measures how far apart the words are and is calculated using the weight (number of 1's) of the XOR function, the number of bits set to one measures the number of bits in which the codewords differ (Fig. 1.8a). The second concept (ii) is the minimum distance, δ , between two codewords and can be defined as Eqn. 1.1.

(a) Hamming distances, ∂

$\begin{array}{r} a = 000 \\ b = 100 \\ \hline \text{XOR} = 100 \\ \partial(a,b) = 1 \end{array}$	$\begin{array}{r} a = 010 \\ b = 100 \\ \hline \text{XOR} = 110 \\ \partial(a,b) = 2 \end{array}$
---	---

(b)

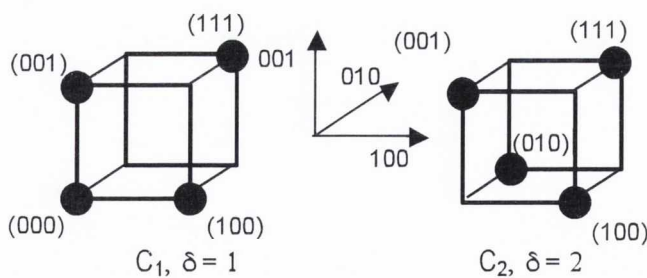


Figure 1.8 (a) illustration of the concept of Hamming distance between codewords, $\partial(a,b)$; (b) codes C_1 , δ (minimum distance) = 1 and C_2 , $\delta = 2$ [33]

$$\delta = \min \{ \partial(a,b) \mid a,b \in C, a \neq b \}$$

Equation 1.1

The larger the minimum distance between two codewords the more digits that must be changed (or errors that must occur) in order to convert one of the words into the other. A simple example of a text message (Table 1.1) can be used to illustrate this point. If an error occurs and message (b) is received then it is immediately clear that an error has occurred, as the message no longer makes sense, and the recipient can ask the sender for clarification. If on the other hand, message (c) were to be received, this is the worst case scenario as this message makes sense but conveys completely the wrong meaning. This example highlights the fact that not all errors are equal. Some are far more serious than others.

Table 1.1 Sample text messages. One error can change a word into another valid codeword.

(a)	I love cake	Intended message
(b)	I love cbke	Non-meaningful error has occurred
(c)	I love jake	Makes sense but gives an incorrect message

Two 3-bit codes are shown in Fig. 1.8b (above), $C_1 = \{000, 001, 100, 111\}$ and $C_2 = \{001, 010, 100, 111\}$. Code C_1 has a minimum distance of 1, meaning that an undetectable error could occur. 000 could be converted into 001 or 100, both valid members of C_1 . C_2 has a minimum distance of two meaning that any one digit error will result in a non-valid codeword. Due to a greater minimum distance, C_2 is a stronger code than C_1 . A simple example to further explain this point can be constructed using a 2-bit code. Let B^n be defined as a binary code composed of 2^n elements, a code C can be defined as $C \subseteq B^n$. For example, B^2 ($B^2 = \{00, 01, 10, 11\}$) could be used as a simple code C_3 to transfer a trivial set of commands (Table 1.2).

Table 1.2 Code C_3

	Up	Down	Left	Right
C_3	00	01	10	11

In this straightforward example of a code our dictionary of codewords is simple and contains only four letters, each two bits long. A sample process for sending a codeword from C_1 can be seen below (Fig. 1.9). First the message is encoded (the desired message is converted using the code dictionary into the binary word designated to represent it) and transmitted. Next, using the same dictionary as before, it is decoded, and in the final step, it

is conveyed to the recipient. If no interference (or noise) occurs during transmission, the message will be delivered error free as intended. If, on the other hand interference interrupts the message during transmission, a distorted codeword could arrive at the designated destination.

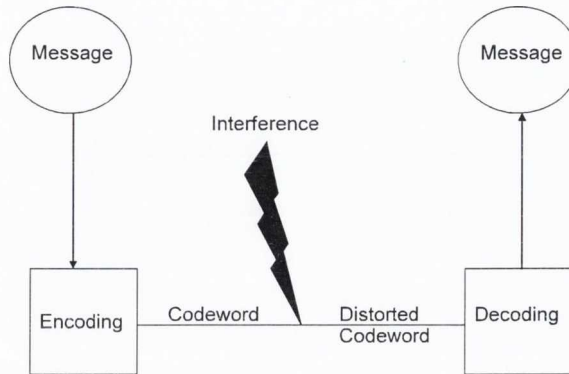


Figure 1.9 Schematic of process for sending a codeword

If a codeword from C_1 01 (down) for example, is sent and an error occurs during transmission resulting in 00 (up) being received, this error would go undetected as 00 (up) is a valid codeword within C_1 . The minimum or limiting distance δ of the code C_1 is 1. A minimum distance of 1 is insufficient to allow for any errors to be detected.

1.2.1 Parity

A useful tool in error detection and correction is the parity of a codeword. A codeword can be described as being of even or odd parity, depending on the number of ones it contains. If the number of ones is even then the word is described as even parity and if the number is odd the word can be described as odd parity. The sample code C_3 discussed above could be improved if an extra bit was added to each codeword to make all the words the same parity, even for example (Fig. 1.10 and Table 1.3).

$$f: B^n \rightarrow B^{n+1}$$

$$f(C) = CX \quad \text{where } X = 0 \text{ if } C \text{ even } X = 1 \text{ if } w \text{ odd}$$

Figure 1.10 Details of function to change all codewords to the same parity

Table 1.3 Code C_4

$C_3=B^2$	$C_4 \subset B^3$
00	000
01	011
10	101
11	110

If one error occurs in any word during the transmission of C_4 , as in example Fig. 1.11, it can be detected, as changing any one bit in a word results in a non-valid codeword. If an error occurs the result will be a non valid codeword.

101----111 (a2)

Figure 1.11 Sample C_4 codeword with one error

The differences seen in the two sample codes indicate that not all codes are in fact equal, some are clearly better than others to use for the transmission of information.

For a more general introduction to error-coding the reader may consult the texts of Biggs and Humphreys [31, 32].

1.3 Nucleotide Donor/Acceptor Patterns

The four bases in DNA form two complementary hydrogen bonded pairs (for further discussion of the hydrogen bond see appendix A1). Recognising the potential importance of hydrogen D/A patterns Mac Dónaill took the approach of assigning each nucleotide a 4-bit binary numerical representation [20] (as introduced in section 1.1)(Fig. 1.3).

Mac Dónaill assigned a pattern to each of the 16 possible nucleotide bases: the structure of each base is as shown in (Fig. 1.2). The 16 molecules (8 complementary pairs) can be divided into two sets based on parity and within these sets a further grouping can be made based on molecular shape (Purine or Pyrimidine). The 16 patterns can be viewed as occupying a hypercube structure (Fig. 1.12). Dividing the pairs by parity yields eight even letters and eight odd, which can be further grouped into 4 even pairs [0000-1111, CG

1001-0110, 0101-1010, 0011-1100] and 4 odd
 [0001-1110, 0111-1000, 1101-0010, 1011-0100].

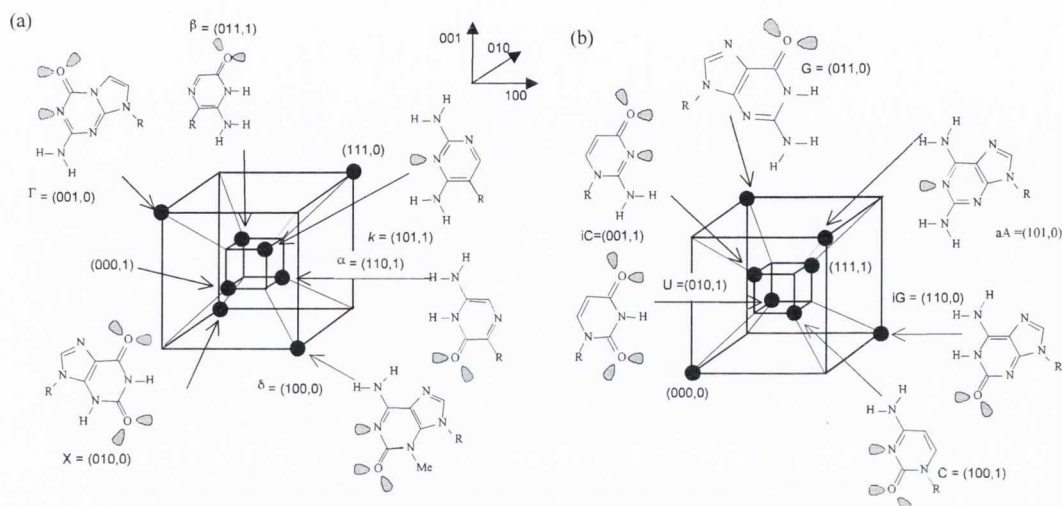


Figure 1.12 Numerical representation of nucleotides depicted as positions on the \mathbf{B}^4 hypercube: (a) odd-parity code; (b) even-parity code. The position occupied by a nucleotide is determined by its D/A pattern. The inner cubes show the “pyrimidines” (final bit = 1), while the outer cubes represent “purines” (final bit = 0).

Diagram taken directly from [21]

If all 16 nucleotides are considered as one large set of molecules, exploring all possible pairings will result in interactions between molecules in which only one mismatch exists (a hydrogen opposing another hydrogen or a lone-pair opposite another lone-pair)(Fig. 1.13). It has been shown that a mismatch in just one out of three D/A positions is not enough to cause repulsion between pairs and a net binding total interaction energy is frequently evident [4]. Thus, an alphabet that allows mismatches of this type will not be resistant to errors.

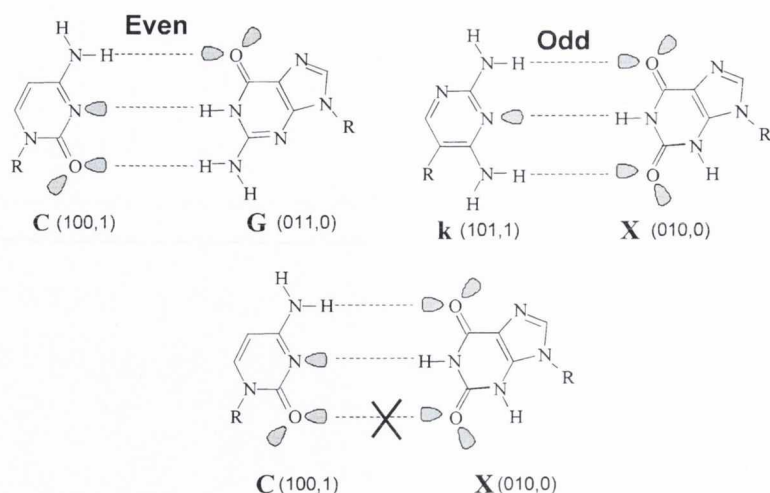


Figure. 1.13 Even parity pair CG and Odd pair $\kappa\chi$. Mixed parity pair $C\chi$ with one mismatch.

The fourth bit of each pattern (assigned based on the purine/pyrimidine structure of each) can be used like a parity check bit to divide the 8 pairs into two separate sets [4].

Alphabets (or subsets of alphabets) containing letters all equal in parity offer a simple but effective error resistance tool. Individual letters are further apart from one another the distance between all letters is 2.

In all cases where pairs are considered in like parity sets, 2 out of a possible 3 D/A positions will mismatch (Fig 1.14, 1.15). This is enough to cause a net repulsive interaction between the two monomers. Error coding theory states that in a code with a minimum distance $\delta=2$ allows for the detection of one error in a given codeword. In the case of nucleotides having a minimum distance of 2 does not allow for an error to be detected but it does increase error-resistance by preventing the formation of non-complementary associations, which with 2 mismatches are energetically non-viable.

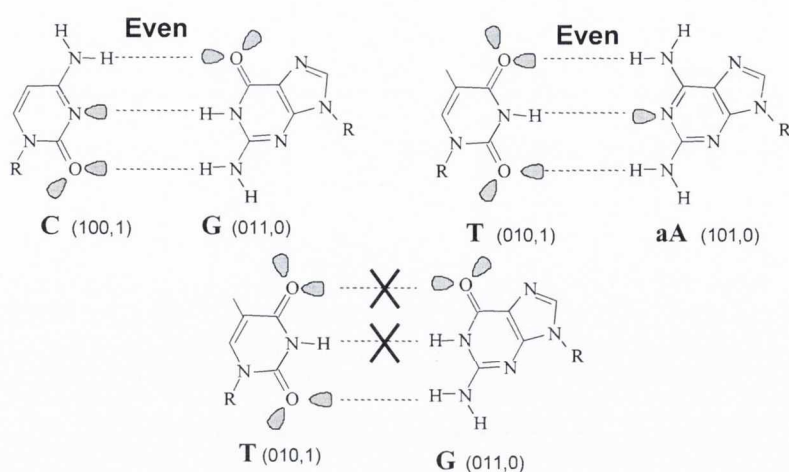


Figure 1.14 Even parity pairs CG TaA. Mixed even parity pair TG with two mismatches.

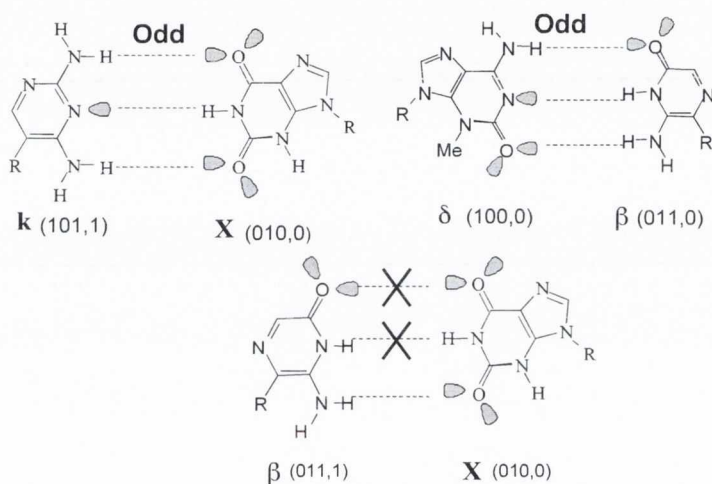


Figure 1.15 Odd parity pairs $\kappa\chi$ $\delta\beta$. Mixed even parity pair $\beta\chi$ with two mismatches.

The overall results (of the quantum chemical calculations and chemical considerations), showed that the largest remaining group is the group found as the basis of our genetic alphabet A (represented in the study by amino-adenine, an idealized form of A), C, G and T [4].

1.4 The advantage of code partitioning

The above analysis offers the first theoretical explanation of the particular composition of the nucleotide alphabet from the set of 16 possible nucleotides or nucleotide analogues. The model, however, does not explain why nucleotides may be preferred over the many conceivable alternatives. This additional question has been more recently been explored by Mac Dónaill who has extended the D/A error coding model beyond nucleotides [33] to

explore the use of two sizes as opposed to one. As this matter is directly pertinent to the problem explored in this thesis we include, with the author's permission, the essential outline of this extended model drawing heavily for a draft manuscript.

In conventional error-coding, where a direct copying process is implicit, a code $C = \{a, b, c, \dots\}$ is employed with few constraints other than the members or codewords must be sufficiently distinct. In a molecular context replication proceeds by template propagation where each codeword a, b, c, \dots , is accompanied by its complement, a^*, b^*, c^*, \dots , and the code may be more accurately described as a set of couples, $C = \{(a,a^*), (b,b^*), (c,c^*), \dots\}$.

This superficially modest additional constraint has surprisingly far reaching consequences for the minimum distance, δ , which is now dependant not only on the various distances, $\partial(a,b)$, as in conventional error-coding, but on the more complicated set of interdependent distances $\partial(a,b)$, $\partial(a,b^*)$, $\partial(a^*,b)$, and $\partial(a^*,b^*)$.

We now consider the more general and abstract system of molecules expressing D/A patterns using n -bits. For two couples (a,a^*) and (b,b^*) , where a and b (and therefore a^* and b^*) differ in r of n positions, then a and b^* , (and a^* and b) will differ in $n-r$ positions (Eqn. 1.2, Eqn 1.3):

$$\partial(a,b) = \partial(a^*,b^*) = r \quad \text{Equation 1.2}$$

$$\partial(a,b^*) = \partial(a^*,b) = n - r \quad \text{Equation 1.3}$$

Combining Eqn 1.2 and Eqn.1.3 we get

$$\partial(a,b) + \partial(a,b^*) = n \quad \text{Equation 1.4}$$

This serves as a significant constraint in the capacity to increase the mutual distinctiveness of codewords since for any increase in $\partial(a,b)$ there is a corresponding decrease in $\partial(a,b^*)$. Thus, reducing the probability of $a \rightarrow b$ errors can only be achieved by increasing the possibility of $a \rightarrow b^*$ errors. The possibility of confusion of a (or a^*) with either b or b^* is least likely where a is simultaneously as dissimilar as possible from both b and b^* , and the

minimum distance for a code composed as few as two complement couples is therefore given by

$$\delta = n/2, n \text{ even} \quad \text{Equation 1.5}$$

$$\delta = (n-1)/2, n \text{ odd} \quad \text{Equation 1.6}$$

Thus, for a code employing n -bits, the maximum mutual separation between codewords is $n/2$, where n is even, and $(n-1)/2$ where n is odd.

The core of the problem in employing D/A patterns alone as the basis of molecular discrimination arises from the inherent conflict in simultaneously attempting to realise sufficiently large values for both $\partial(a,b)$ and $\partial(a,b^*)$, where $\partial(a,b)$ can be increased only at the expense of decreasing $\partial(a,b^*)$. Codes composed of complement couples in which the elements are distinguished only by their D/A patterns as in C_A (Fig. 1.16), are error-prone since the maximum value for the ‘minimum distance’ is capped at $\delta = n/2$, n even, and $\delta = (n-1)/2$, where n is odd.

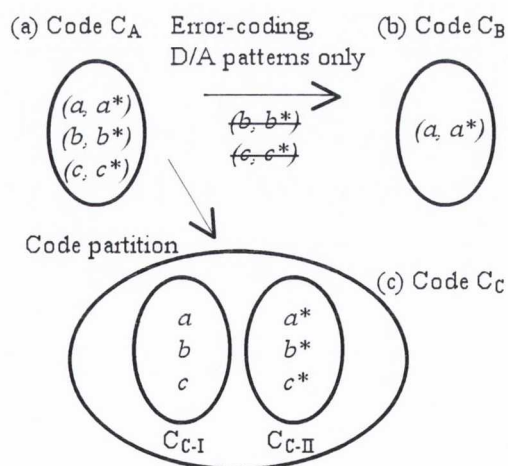


Figure 1.16 (a) Code C_A , a code composed of complement couples; (b) C_B , reduced to a single complement couple, and; (c) Code C_C partitioned into two subcodes.

Fig. 1.16(c) depicts a code, C_C , similar to C_A but partitioned into two subcodes, C_{C-I} and C_{C-II} , according to some suitable physicochemical molecular property, such that where a codeword, say a , is a member of one subcode, its complement, a^* , will be a member of the

other. One imagines that this distinguishing feature of the subgroups can be harnessed so as to ensure that physical association only occurs between members of C_{C-I} and members of C_{C-II} but that associations within subgroups are prohibited. The role of Hamming distances between D/A patterns can now be reduced to optimising the mutual distinguishability of members within one subgroup, e.g. C_{C-I} , and, since there is a one to one mapping between the membership of C_{C-I} and that of C_{C-II} , it follows that the mutual Hamming distances between members of C_{C-I} will precisely mirror those between members of C_{C-II} , i.e. $\partial(a,b) = \partial(a^*,b^*)$ for any a and b .

Since membership of C_{C-I} or C_{C-II} is now determined by a non-D/A feature, Hamming distances between members of C_{C-I} and members of C_{C-II} , that is, distances of the type $\partial(a,b^*)$ or $\partial(a^*,b)$, cease to be of direct relevance. Consequently, constraints of the type expressed in Eqn.1.4 no longer limit the maximum mutual distance since $\partial(a,b)$ may be increased as required without regard to the consequences for $\partial(a,b^*)$.

All that now matters in regard to D/A patterns are the Hamming distances $\partial(a,b)$, $\partial(a,c)$ and $\partial(b,c)$. The extent to which $a \in C_{C-I}$, for example, can incorrectly associate with the reduced set of available non-complements, b^* and c^* , in C_{C-II} depends on their dissimilarity to the intended complement, a^* , i.e. $\partial(a^*,b^*)$ and $\partial(a^*,c^*)$; considering all members of C_{C-I} , a , b , and c , the relevant Hamming distances are $\partial(a^*,b^*)$, $\partial(a^*,c^*)$ and $\partial(b^*,c^*)$. Similarly, the non-complementary associations of a^* , b^* , and $c^* \in C_{C-II}$ with members of C_{C-I} depend on $\partial(a,b)$, $\partial(a,c)$ and $\partial(b,c)$. These two distance sets are equivalent since distances between corresponding codewords are preserved on moving between subcodes, i.e. $\partial(a,b) = \partial(a^*,b^*)$. Most significantly, as D/A patterns are now charged only with distinguishing between codewords within the same subcode, the debilitating conflict in simultaneously addressing $\partial(a,b)$ and $\partial(a,b^*)$ is removed. The preferred combination of D/A patterns within a subcode may be determined in precisely the same manner as in conventional error-detecting codes, allowing codewords to be as dissimilar as necessary, without conflicting pressures.

There are three primary concepts in relating molecular codes to computer binary codes:

Computer codes (sets of codewords) work by a one step copying process where a text is directly copied codeword by codeword. Complements are not required and the mutual distinctiveness of codewords may be made as large necessary.

- (i) In a molecular context replication proceeds by template propagation where a replication of an original text requires the creation of a negative text. The replication procedure must be executed twice to produce a copy of the original. This two-step process requires that each codeword in a code be accompanied by the complementary pattern, and this as we have seen greatly limits the maximum available mutual distinctiveness of codewords.
- (ii) Partitioning a code into two subsets based on some feature other than D/A patterns precisely offsets the constraint in (i) above and removes the cap on mutual distinctiveness.

In the nucleotide alphabet this non D/A feature is the purine/pyrimidine size asymmetry. Thus, the two sizes in nucleotide alphabet may be directly related to the two step replication process which in turn arises from template rather than direct-copy replication.

1.5 Designing an alternative to nucleotides

In nucleotide bases the first 3 bits of each numerical representation come directly from the hydrogen D/A pattern and the fourth comes from the size of the molecule. In order to study D/A patterns further and explore the significance of how the molecular information is displayed, a reverse engineering approach will be undertaken. In this approach a set of molecules will be designed that will have the same 4-bit numerical representations as the bases in our terrestrial genetic alphabet, but these molecules used to represent the information (D/A pattern) will have different structures compared to conventional nucleotides. The area of supramolecular chemistry provides many examples of the types of hydrogen bonding systems that could be considered, particularly the work of Lehn [6]. Literature provides some examples of molecules capable of forming quadruple hydrogen bonds [5][34]Fig. 1.17).

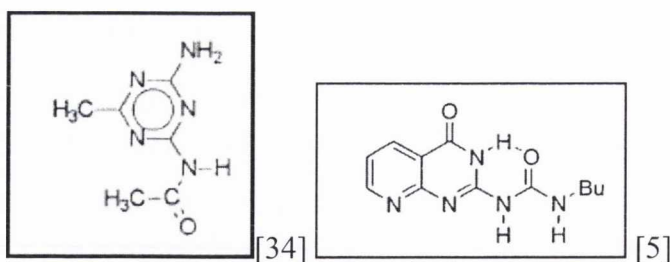


Figure 1.17 Molecule capable of forming quadruple hydrogen bonds Pictures taken directly from [34][5]

Before considering constructing a set of molecules from literature an ideal set will be constructed and modelled. This ideal set of molecules will be designed in such a way that all things are kept as uniform as possible throughout the set. With this uniformity in mind, and due to its rigid structure and shape that can easily accommodate 4 D/A sites, a naphthalene structure will be used as the basic template for each molecule (Fig. 1.18).

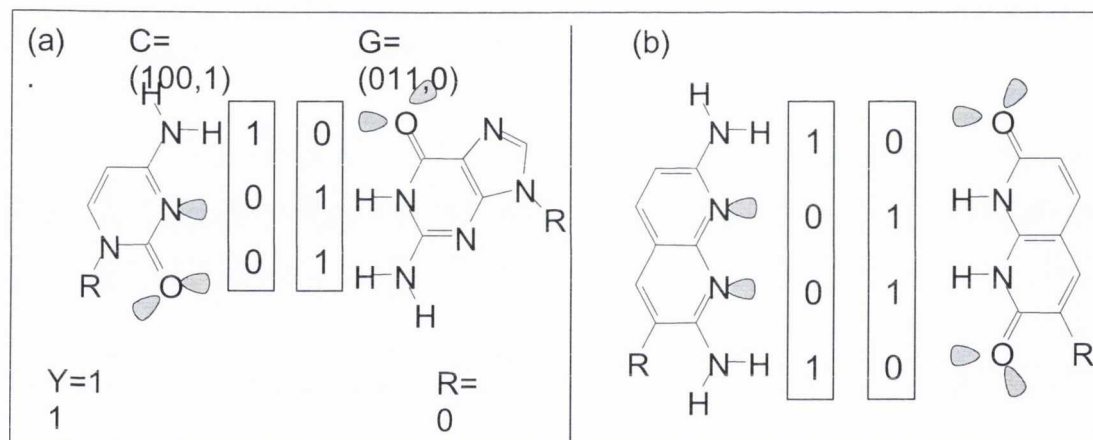


Figure 1.18 (a) Nucleotide pair CG (b) Informationally equivalent heteronaphthalene pair

As the desired outcome is to explore the entire set of possible 4 D/A heteronaphthalene (collectively termed Het) structures, 16 molecules in total must be constructed (2^4), one for each unique D/A pattern (Fig. 1.19). Full details of the construction of the Het set of molecules can be seen in 3.1. In order for a viable alphabet to be formed from the set of Het molecules (Table 1.4);

Table 1.4 Table of viability requirements

1	Each molecule must bind to its complementary molecule
2	Each molecule should repel any molecule with which it does not form a complementary pair.
3	Any surviving molecules must comply with chemical constraints (although this will not be a primary concern in this thesis).

To determine if the results gathered from the study of the proposed ideal Het potential alphabet a further study will be carried out based on a set of molecules constructed (where possible) from literature. This “real” set of molecules was primarily based on the work of Zimmerman and Corbin [5] (Fig. 1.20). Full details on building this set of molecules can be found in 7.1. The results of the Zimmerman (Zim) alphabet will be compared to the Het.

Before exploring the Het and Zim potential alphabet letter sets it is important to consider the computational methods that will be used. With this in mind the next chapter will detail the methods used in this thesis.

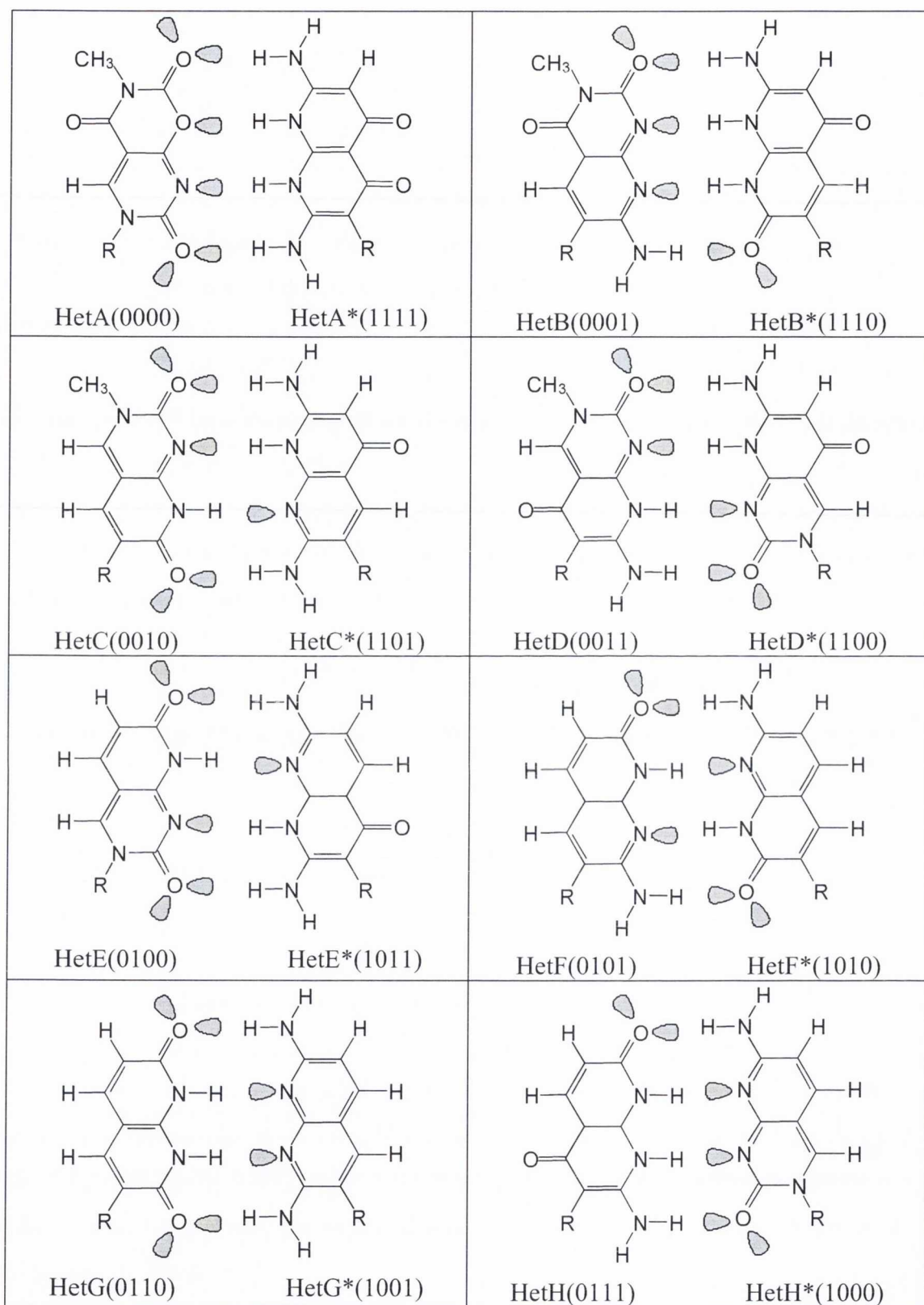


Figure. 1.19 The 16 possible 4-bit hydrogen/lone pair patterns expressed in 'heteronaphthalenes'

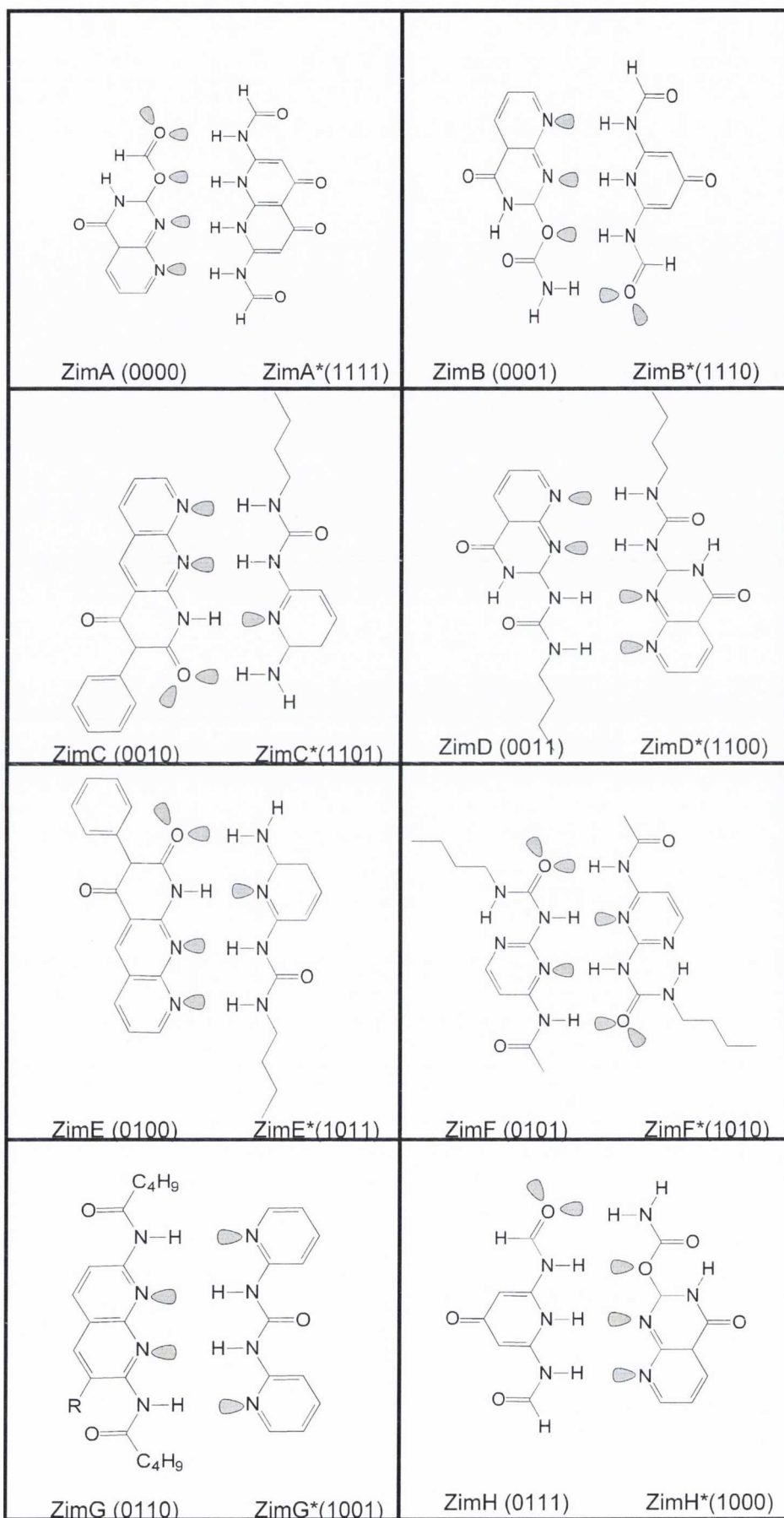


Figure. 1.20 Zimmerman alphabet. 16 letters/molecules broken into 8 complementary pair

1. Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids- A structure for deoxyribose nucleic acid*. Nature, 1953. **171**(4356): p. 737-738.
2. Switzer, C., S.E. Moroney, and S.A. Benner, *Enzymatic incorporation of a new base pair into DNA and RNA*. Journal of the American Chemical Society, 1989. **111**(21): p. 8322-8323.
3. Piccirilli, J.A., et al., *Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet*. Nature, 1990. **343**(6253): p. 33-37
4. MacDonaill, D.A. and D. Brocklebank, *An ab initio quantum chemical investigation of the error-coding model of nucleotide alphabet composition*. Molecular Physics, 2003. **101**(17): p. 2755-2762.
5. Sijbesma, R.P. and E.W. Meijer, *Quadruple hydrogen bonded systems*. Chemical Communications, 2003(1): p. 5-16.
6. Zimmerman, S.C. and F.S. Corbin, *Heteroaromatic modules for self-assembly using multiple hydrogen bonds*. Molecular Self-Assembly, 2000. **96**: p. 63-94.
7. Lehn, J., Marie, *Supramolecular Chemistry, Concepts and Perspectives* 1995: VCH
8. Einstein, A., in *Festschrift fur Aurel Stodola*, . E. Honegger, Ed. Vol. Orell Fussli Verlag, Zurich. 1929
Quote as translated by
Eschenmoser, A., *Chemical etiology of nucleic acid structure*. Science, 1999. **284**(5423): p. 2118-2124.
9. Miller, S.L., *A Production of amino acids under possible primitive earth conditions*. Science, 1953. **117**(3046): p. 528-529.
10. Wachtershauser, G., *Evolution of the first metabolic cycles*. Proceedings of the National Academy of Sciences of the United States of America, 1990. **87** : P. 200-204.
11. Szathmary, E., *4 Letters in the genetic alphabet- A frozen evolutionary optimum*. Proceedings of the Royal Society of London Series B-Biological Sciences, 1991. **245**(1313): p. 91-99.
12. Evans, T.A. and K.R. Seddon, *Hydrogen bonding in DNA - a return to the status quo*. Chemical Communications, 1997(21): p. 2023-2024.
13. Roberts, C., R. Bandaru, and C. Switzer, *Theoretical and experimental study of isoguanine and isocytosine: Base pairing in an expanded genetic system*. Journal of the American Chemical Society, 1997. **119**(20): p. 4640-4649.
14. Guckian, K.M., T.R. Krugh, and E.T. Kool, *Solution structure of a nonpolar, non-hydrogen-bonded base pair surrogate in DNA*. Journal of the American Chemical Society, 2000. **122**(29): p. 6841-6847.
15. Kool, E.T., J.C. Morales, and K.M. Guckian, *Mimicking the structure and function of DNA: Insights into DNA stability and replication*. Angewandte Chemie-International Edition, 2000. **39**(6): p. 990-1009.
16. Liang, W., et al., *Systematic theoretical investigations on all of the tautomers of guanine: From both dynamics and thermodynamics viewpoint*. Chemical Physics, 2006. **328**(1-3): p. 93-102.
17. Dawkins, R., *The Blind Watchmaker*. Originally published by Longman Scientific & Technical 1986, 1991: p. 112.
18. Yockey, H.P., *Information Theory and Molecular Biology*. CUP, 1992: p. 102.
19. Szathmary, E., *What is the optimum size for the genetic alphabet*. Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(7): p. 2614-2618.
20. Mac Donaill, D.A., *A parity code interpretation of nucleotide alphabet composition*. Chemical Communications, 2002(18): p. 2062-2063.

21. Mac Donaill, D.A., *Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition*. *Origins of Life and Evolution of the Biosphere*, 2003. **33**(4-5): p. 433-455.
22. Sinden, R.R., *DNA structure and function*. 1994: Academic Press
23. Neidle, S., *Nucleic acid structure and recognition*. OUP, 2002.
24. Sinden, R.R., *DNA structure and function*. 1994: Academic Press p.10
25. Dickerson, R.E., et al., *The anatomy of A-DNA, B-DNA, and Z-DNA*. *Science*, 1982. **216**(4545): p. 475-485.
26. Sinden, R.R., *DNA structure and function*. 1994: Academic Press p.23
27. Eschenmoser, A., *Chemical etiology of nucleic acid structure*. *Science*, 1999. **284**(5423): p. 2118-2124.
28. Eschenmoser, A., *Hexose Nucleic-Acid*. *Pure and Applied Chemistry*, 1993. **65**(6): p. 1179-1188.
29. Eschenmoser, A. and M. Dobler, *Why pentose and not hexose nucleic-acids. 1. Introduction to the problem, conformational-analysis of oligonucleotide single strands containing 2',3'-dideoxycyclopentanosyl building-blocks(homo-DNA), and reflections on the conformation of A-DNA and B-DNA*. *Helvetica Chimica Acta*, 1992. **75**(1): p. 218-259.
30. Hamming, R.W., *Error detecting and error correcting codes..* *Bell System Technical Journal*, 1950. **29**(2): p. 147-160. 30.
31. Biggs, N., *Discrete Mathematics (second edition)*. OUP, 2002.
32. Humphreys, J.F., Prest, M.Y., *Numbers, Groups & Codes (second edition)*. CUP, 2004.
33. Mac Donaill, D.A., *Molecular Error-Coding: Why Nucleotides Come in Two Sizes In preparation*.
34. Beijer, F.H., et al., *Self-complementarity achieved through quadruple hydrogen bonding*. *Angewandte Chemie-International Edition*, 1998. **37**(1-2): p. 75-78.

2 Computational Theory and Methods

2.1 Introduction

Computational techniques provide molecular information and also provide data on molecule systems which are often impractical or infeasible to determine experimentally. As the speed and power of computers has grown, so has the quantity and power of computational methods available. A large variety of software exists for use in the determination of chemical information. In this thesis two packages were used, Gaussian 03W and Spartan 04 V1.0.1. Gaussian 03W was used in conjunction with the Trinity Centre for High Performance Computing (TCHPC) IITAC project clusters. Both GaussView (the viewer available with Gaussian03W) and Spartan were run using a desktop PC.

Although computational methods have increased in number and complexity, the most appropriate choice may not always be the most expensive method available. The use of the result needs to be considered as well as weighed against the computational expense of the calculation. This time versus accuracy trade-off was kept in mind when deciding on the most appropriate methods to use in this work. In the study of molecular dimers undertaken in this thesis, having accurate individual results that are in close agreement with experimental results is not the primary concern. The focus here is on determining the overall relative pattern of results in a consistent fashion that allows direct comparison of different data sets. With these things in mind it was decided to use Hartree-Fock (HF) with a 6-31G* basis set as the initial calculation method for the exploration of molecular interaction energies. Although using HF will not give rise to results as close to experimental values as possible, it will still give realistic results sufficient for the purpose of overall pattern determination and data set comparison required in this work [1].

2.1.2 Schrödinger Equation

In writing this discussion of computational theory and methods [2] was used. The Schrödinger equation lies at the centre of quantum mechanics. The time independent

version of the Schrödinger equation is Eqn. 2.1a, where \hat{H} is the Hamiltonian operator, E is the energy and Ψ is the wavefunction. The Hamiltonian operator can be written as Eqn. 2.1b for n particles, where $\hbar = h/2\pi$ (h is Planck's constant), m is the mass of the particle, i labels the particle, V represents the potential energy and ∇_i^2 the kinetic energy operator is given by a Laplacian (Eqn. 2.1c).

$$\hat{H}\Psi = E\Psi \quad \text{Equation 2.1a}$$

$$\hat{H} = -\frac{\hbar^2}{2} \sum_{i=1}^n \frac{1}{m_i} \nabla_i^2 + V(r) \quad \text{Equation 2.1b}$$

$$\nabla_i^2 = \left\{ \frac{\delta^2}{\delta x_i^2} + \frac{\delta^2}{\delta y_i^2} + \frac{\delta^2}{\delta z_i^2} \right\} \quad \text{Equation 2.1c}$$

The Schrödinger equation can only be solved exactly for one electron systems, such as H and He^+ ; it is impossible to solve it exactly for any system with more than one electron due to electron-electron interactions present. The infeasibility of solving a three or more body system implies that any solution can only be an approximation of the exact answer.

2.1.3 Variation Theorem

As there is no exact numerical solution to the Schrödinger equation for a multi body system, some method of determining which approximate solution is the best is required. The variation theorem provides such a method, it states that the energy determined using an approximation of the exact wavefunction will always be greater than the energy determined with the true wavefunction. Thus the value for the trial wavefunction can be varied and the best determined to be that with the lowest energy.

2.1.4 Born Oppenheimer Approximation

To aid in solving the Schrödinger equation the wavefunction can be split into two parts, nuclear and electronic. This simplification is possible because the movement of nuclei

(due to large mass) is much slower than the movement of electrons. Under the Born Oppenheimer approximation the nuclei have zero kinetic energy as they are seen as fixed in space when compared to the motion of fast moving electrons. The potential energy of the nucleus-nucleus interactions must still be considered but it needs only to be calculated once for each atomic configuration. With these simplifications in place the electronic Schrödinger equation is made up of just three terms (Eqn 2.2). The first term is the kinetic energy of the electron, the penultimate is the potential energy of the electron-nucleus (eN) interaction and the final term is the potential energy of the electron-electron (e) interactions.

$$H_e = K_e + V_{eN} + V_e \quad \text{Equation 2.2}$$

2.2 Hartree-Fock Method

A multielectron problem is impossible to solve exactly (as discussed above). A Self-consistent field (SCF) procedure developed by Hartree [3] simplified the problem by neglecting individual electron-electron interactions and assuming that electrons move in an averaged field created by all other electrons and nuclei. The Hartree product is used to represent the total wavefunction as a product of individual one electron wavefunctions (Eqn. 2.3).

$$\Psi^{HP} = \chi_i(x_1)\chi_j(x_2)\chi_k(x_3)\dots\chi_n(x_N) \quad \text{Equation. 2.3}$$

In the SCF procedure an initial guess at a set of orbitals is made and then used to solve the Schrödinger equation and generate a new set of orbitals. This new set of orbitals is then used to start the process again and the procedure is repeated until the total energies have converged. This method was improved on by Fock[4] who addressed Hartree's neglect of the Pauli principle, which states that upon exchange of two electrons a change should be seen in the sign of the wavefunction. The anti symmetric wavefunction is taken into account through a single Slater determinant (Eqn. 2.4). With the inclusion of the Pauli principle to the Hartree model the improved Hartree-Fock model was formed.

$$\Psi = \frac{1}{\sqrt{N}} \begin{vmatrix} \chi_i(x_1) \dots & \dots \chi_j(x_1) \\ \chi_i(x_N) \dots & \dots \chi_j(x_N) \end{vmatrix} \quad \text{Equation 2.4}$$

In order to find the HF energy the wavefunction is expanded by introducing a linear combination of atomic orbitals (LCAO) (Eqn. 2.5).

$$\chi_i = \sum_{\mu} c_{\mu}^i \phi_{\mu} \quad \text{Equation 2.5}$$

The molecular orbital is given by Ψ_i , the atomic orbital by ϕ_{μ} and the coefficient for the atomic orbital ϕ_{μ} in the molecular orbital Ψ_i is given by c_{μ}^i .

In order to find the Hartree-Fock energy the Roothaan-Hall equation is utilised to rewrite the wavefunction (Eqn 2.6). F is the Fock matrix, C is the coefficient matrix defined by LCAO, S is the overlap matrix and \mathcal{E} are the orbital energies.

$$FC = SCE \quad \text{Equation 2.6}$$

Calculation of the Hartree-Fock energy results in two-electron integrals for the Coulomb (Eqn. 2.7a) and Exchange (2.7b). These two-electron integrals carry the largest computational cost during a HF calculation; formally they scale as M^4 in which M is the number of basis functions.

$$J_{ij} = \int \int \chi_i(1) \chi_j(2) \frac{1}{r_{12}} \chi_i(1) \chi_j(2) d\tau_1 d\tau_2 \quad \text{Equation 2.7a}$$

$$K_{ij} = \int \int \chi_i(1) \chi_j(2) \frac{1}{r_{12}} \chi_i(2) \chi_j(1) d\tau_1 d\tau_2 \quad \text{Equation 2.7b}$$

A major flaw with the HF method is its neglect of correlation. Although the variation theorem ensures that the lowest energy trial solution E_τ is chosen there is no way of telling how great the distance is from this to the true energy E_0 . The difference between these values is described as the correlation energy. The main component of the correlation energy arises from the fact that electrons do not move completely independently of each other. Where possible they move in ways to avoid each other and thus to minimise the build up of charge. Correlation can be included through the use of post-HF methods namely Configuration Interaction (CI) or Møller-Plesset (MP). In this thesis MP2 (second order Møller-Plesset) will be considered. Post-HF methods add correlation through mixing excited-state and ground-state wavefunctions, extending the flexibility of the HF model. In this thesis MP2 will be used to explore the effect of correlation.

2.3 Basis Set Choice

A basis set is a mathematical representation of molecular orbitals which is formed through a linear combination of basis functions. In *ab initio* calculations such as HF, Gaussian functions (Eqn. 2.8) are used in the generation of the wavefunction. A linear combination of primitive Gaussians is used to form a contracted Gaussian.

$$G(\alpha, r) = Ae^{-\alpha r^2} \quad \text{Equation 2.8}$$

A popular choice of basis set and one used in this work is a split valence basis set. In a split valence basis set functions are split according to core electrons, inner valence electrons and outer valence electrons. For example in a 6-31G basis, 6 contracted Gaussians are used for the core electrons, 3 contracted Gaussians for the inner valence electrons and one Gaussian function for the outer valence. In order to improve the description of valence electrons using a Gaussian basis set, polarisation can be taken into account. Polarisation arises as a consequence of the charge clouds of two atoms on bonding being directed in a particular spatial direction. Adding a polarisation function (shown by the letter of the orbital being added or often * as in the original Pople notation) essentially adds one extra unit of angular momentum than the highest occupied unit of angular momentum. Let us consider methane (CH_4) if one degree of polarisation is added

as part of a basis set choice, one set of p functions would be added for each hydrogen atom and one set of d functions for the carbon. Diffuse functions can also be added to basis sets (represented by a +), the most common use for these functions is for excited states. Test calculation results and further discussion of basis set choice can be found in appendix A3.

2.4 Basis Set Superposition Error (BSSE)

When calculating the interaction energy of a molecular complex AB, a discrepancy arises due to the inconsistency in basis set size between the individual components A and B and the complex AB. The basis set description of A is improved by using the basis functions of B and vice versa. This use of extra basis functions (for each monomer) results in an artificially large and flexible basis set. The error created in the interaction energy is referred to as basis set superposition error (BSSE). BSSE results in the overestimation of intermolecular interactions. In this thesis we shall focus on the calculation of the total interaction energy (TIE - the difference between monomers and dimer when the monomers have geometry independent of the dimer)[‡] of the AB complex. The TIE is calculated by Eqn. 2.9[†], in which each structure (monomer or dimer) has the form $E_Y^Z(X)$, where $E(X)$ equals the energy of X (X= A, B or AB), Y is the geometrical arrangement of X (Y = A, B or AB) and Z is the basis set of X.

$$E_{\text{TIE}}(\text{AB}) = E_{\text{AB}}^{\text{AB}}(\text{AB}) - (E_{\text{A}}^{\text{A}}(\text{A}) + E_{\text{B}}^{\text{B}}(\text{B})) \quad \text{Equation 2.9}$$

The first popular way of removing BSSE, known as counterpoise (CP), was proposed by Boys & Bernardi [5] in 1970. In their method the CP correction is added to each monomer by using all of the basis functions of the dimer, thus making monomers and dimers directly comparable. The Boys & Bernardi method was thought to overcorrect for BSSE [6,7].

[‡] Total interaction energy as used here is often referred to as the binding energy. The binding energy would formally be equal in magnitude to the Total interaction energy as defined above in Eqn. 2.9 but opposite in sign.

[†] The notation used is adapted from [8].

A new counterpoise scheme was proposed by Simon, Duran & Dannenberg [9], in which BSSE can be taken into account during optimisation of the dimer thus correcting the potential energy surface and minimum energy determined for the supermolecular structure.

Alternative superposition error correction methods have been put forward in the literature [10, 11, 12], although the Boys & Bernardi method is still widely used [13, 14], particularly in conjunction with the Gaussian programme.

Although there is no doubt that BSSE will affect the overall binding strength of a molecular dimer, some authors have argued that attempting to remove it does not necessarily lead to a more accurate result, and the largest basis set possible should simply be used [15]. In other cases it can be argued that its removal is not strictly necessary. For example if comparing like systems the error will be roughly equal across the systems resulting in only a small change in the overall result on its removal [16].

To overcome the problem of BSSE several approaches can be used in conjunction with Gaussian 03W;

1. The basis set of each monomer can be made consistent with that of the dimer as in Eqn. 2.10.

$$E_{IE}(AB) = E_{AB}^{AB}(AB) - (E_{AB}^{A+ghost\ B}(A) + E_{AB}^{B+ghost\ A}(B)) \quad \text{Equation 2.10}$$

This method can only be used when calculating the interaction energy (IE - energy difference between the monomers and the dimer when the monomers have the dimer geometry). In this method when calculating the energy of molecule A, ghost atoms (atoms of a specified type with normal basis functions having no electrons or nuclear charge) are used to represent the basis set of molecule B thus creating a basis set comparable to that of the dimer AB.

2. CP can be used on the dimer structure to correct for BSSE. This method takes into account the number of molecules or molecular fragments present and uses this to correct for the over estimation in basis set size. Counterpoise can be added in one of two ways using Gaussian 03W:

2A. After optimisation during the calculation of single point energy (Eqn. 2.11a, Eqn. 2.11b).

$$E_{IE}(AB) = E_{AB}^{AB-CP}(AB) - (E_A^A(A) + E_B^B(B)) \quad \text{Equation 2.11a}$$

$$E_{TIE}(AB) = E_{AB}^{AB-CP}(AB) - (E_A^A(A) + E_B^B(B)) \quad \text{Equation 2.11b}$$

2B. During the geometry optimisation (Eqn. 2.12a) (Eqn. 2.12b).

$$E_{IE}(AB) = E_{ABCP}^{AB-CP}(AB) - (E_A^A(A) + E_B^B(B)) \quad \text{Equation 2.12a}$$

$$E_{TIE}(AB) = E_{ABCP}^{AB-CP}(AB) - (E_A^A(A) + E_B^B(B)) \quad \text{Equation 2.12b}$$

In order to explore the effect of BSSE on TIEs and to gain insight into how the magnitude of the error changes using different calculation techniques, the three available approaches are considered, all using the Duran & Dannenberg [9] counterpoise calculation method.

Path 1. Total neglect of BSSE

Path 2. CP included after optimisation

Path 3. CP included during optimisation

It was decided that for mismatched pairs BSSE would be taken into account using path 2. This allows the results from paths 1 and 2 to be compared and the superposition error to be given a value for each molecular pair. Sample calculations for the comparison of the three paths can be seen in appendix A2.

2.5 Semi-empirical Methods

Semi-empirical methods can offer a fast alternative to performing *Ab Initio* calculations such as Hartree-Fock (HF). HF calculations scale rapidly (formally M^4 where M is the number of basis functions) and calculations become costly time wise even for quite small molecules. The core simplification used in semi-empirical methods is Zero Differential

Overlap (ZDO). This simplification means that any overlap between atomic orbitals situated on different atomic centres is disregarded. Neglecting overlap in this way has several consequences;

- **S** the overlap matrix becomes the identity matrix.
- Integrals for one-electron three centres are set to zero.
- Integrals for two-electron three and four centres are omitted.

Starting from the use of ZDO semi-empirical methods have evolved, leading to the two well known methods AM1 and PM3 considered in this work (Table 2.1).

Table 2.1 Summary of Semi-empirical methods evolution.

NDDO (Neglect of Diatomic Differential Overlap) ZDO
INDO (Intermediate Neglect of Differential Overlap) ZDO Further neglect of two-electron two centre integrals (all except Coulomb type)
CNDO (Complete Neglect of Differential overlap) ZDO All two-electron integrals (including those remaining in INDO) are approximated
MINDO/3 (Modified Intermediate Neglect of Differential Overlap) Derived from INDO, parameterisation added based on experimental data added for some elements.
MNDO (Modified Neglect of Diatomic Overlap) Derived from NDDO Only valence s and p functions considered Parameterization based on atomic spectra and fitting to molecular data

2.5.1 AM1 (Austin Model 1) and PM3 (Modified Neglect of Diatomic Overlap, Parametric Method 3)

AM1 [17] was developed in the laboratory of Dewar in 1985. It is based on MNDO but attempts to combat problems that arose due to the repulsion between atoms being overestimated. Core-core functions were modified to fix this problem.

PM3 is the third parameterisation of MNDO. It was proposed by J. J. P. Stewart in 1989[18]. It differs from MNDO and AM1 in that the parameterisation is automated.

Both AM1 and PM3 suffer from a number of limitations. Those relating to the work in this thesis include[19];

- Incorrect hydrogen bond geometry is often predicted with the use of AM1 (although in strength they are approximately correct)
- The stability of alkyl groups is overestimated in AM1
- With PM3 hydrogen bonds are too short
- In contrast to what is seen experimentally, sp^3 nitrogen atoms are predicted to be pyramidal when using PM3.
- In general weak interactions such as hydrogen bonds can be badly predicted.

These limitations will be kept in mind when considering and comparing results found using these semi-empirical methods.

2.6 Møller-Plesset Model

One way in which electron correlation can be included is the use of a post-HF method namely Configuration Interaction (CI) or Møller-Plesset (MP[20]). In this work MP2 (second order Møller-Plesset) will be considered. Post-HF methods add correlation through mixing excited-state and ground-state wavefunctions, extending the flexibility of the HF model.

MP2 is based upon perturbation theory. This theory centres on the idea that a complex problem can be broken into parts, one part of which has a solution that is known and the

other part has a value which is close to that of the known differing by only a small amount. In the context of MP2 we know the HF Hamiltonian and can use this as a starting point to find the exact Hamiltonian H by adding a perturbation $H^{(0)}$ (Eqn. 2.13).

$$H = H^{(0)} + \lambda v \quad \text{Equation 2.13}$$

The exact wavefunction (Ψ) and Energy (E) can be expanded in terms of the HF wavefunction and energy (Eqn. 2.14a, 2.14b)

$$E = E^{(0)} + \lambda E^{(1)} + \lambda^2 E^{(2)} + \lambda^3 E^{(3)} \dots \quad \text{Equation 2.14a}$$

$$\Psi = \Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \lambda^3 \Psi^{(3)} \dots \quad \text{Equation 2.14b}$$

These equations can then be substituted into the Schrödinger equation, expanded and grouped together in terms of power of the perturbation parameter (Eqn. 2.15a, 2.15b, 2.15c).

$$H \Psi = E \Psi \quad \text{Equation 2.15a}$$

$$H \Psi^{(1)} + v \Psi^{(0)} = E^{(0)} \Psi^{(1)} + E^{(1)} \Psi^{(0)} \quad \text{Equation 2.15b}$$

$$H \Psi^{(2)} + v \Psi^{(1)} = E^{(0)} \Psi^{(2)} + E^{(1)} \Psi^{(1)} + E^{(2)} \Psi^{(0)} \quad \text{Equation 2.15c}$$

The first order correction to the energy is given by $E^{(1)}$, the second by $E^{(2)}$ etc. These can be calculated by the following integrals (Eqn. 2.16a, 2.16b)

$$E^{(1)} = \int \Psi^{(0)*} v \Psi^{(0)} d\tau \quad \text{Equation 2.16a}$$

$$E^{(2)} = \int \Psi^{(0)*} v \Psi^{(1)} d\tau \quad \text{Equation 2.16b}$$

In order to determine the energy corrections for a given order the wavefunction for that order must also be calculated. Adding the zeroth and first- order energies gives the Hartree-Fock energy. In order to better HF the use second or higher order Moller-Plassett perturbation theory is required. The correlation energy can be expressed as the addition of second order corrections and above (Eqn. 2.16b)(For full derivation see [2]).

By using a series of different calculation methods from semi-empirical and *ab initio* through to MP2, which includes correlation, it is hoped to rule out any artefacts which may be present in any results set, due purely to the choice of calculation method.

1. Guo, H. and M. Karplus, *Ab Initio studies of polyenes.1. 1,3-butadiene*. Journal of Chemical Physics, 1991. **94**(5): p. 3679-3699.
2. Leach, A.R., *Molecular Modelling Principles and Applications (second edition)*. Prentice Hall, 2001
3. Hartree, D., *The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods*. Proceedings of the Cambridge Philosophical Society, 1928. **28**.
4. Fock, V., *Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems*. Zeitschrift fuer Physik, 1930. **62**.
5. Boys, S.F. and F. Bernardi, Calculation of small molecular interactions by differences of separate total energies - some procedures with reduced errors. Molecular Physics, 1970. **19**(4): p. 553-&.
6. Johansson, A., P. Kollman, and Rothenberg, S., *Application of functional Boys Bernardi counterpoise method to molecular potential surfaces*. Theoretica Chimica Acta, 1973. **29**(2): p. 167-172
7. Frisch, M.J., et al., *Extensive theoretical-studies of the hydrogen-bonded complexes (H₂O)₂, (H₂O)₂H⁺, (HF)₂, (HF)₂H⁺, F₂H⁻, AND (NH₃)₂*. Journal of Chemical Physics, 1986. **84**(4): p. 2279-2289.
8. Salvador, P., *Implementation and application of basis set superposition error-correction schemes to the theoretical modeling of weak intermolecular interactions*. Doctoral thesis 2001: Department of Chemistry and Institute of Computational Chemistry, University of Girona.
9. Simon, S., M. Duran, and J.J. Dannenberg, *How does basis set superposition error change the potential surfaces for hydrogen bonded dimers?* Journal of Chemical Physics, 1996. **105**(24): p. 11024-11031.
10. Salvador, P., D. Asturiol, and I. Mayer, *A general efficient implementation of the BSSE-free SCF and MP2 methods based on the Chemical Hamiltonian Approach*. Journal of Computational Chemistry, 2006. **27**(13): p. 1505-1516.
11. Galano, A. and J.R. Alvarez-Idaboy, *A new approach to counterpoise correction to BSSE*. Journal of Computational Chemistry, 2006. **27**(11): p. 1203-1210.
12. Sadlej, A.J., *Exact perturbation treatment of the basis set superposition correction*. Journal of Chemical Physics, 1991. **95**(9): p. 6705-6711.
13. Mirzaei, M. and N.L. Hadipour, *A computational NQR study on the hydrogen-bonded lattice of cytosine-5-acetic acid*. Journal of Computational Chemistry, 2008. **29**(5): p. 832-838.
14. Zheng, X.Y., et al., *Density Functional Theory Study of the Free and Tetraprotonated Spheroidal Macrotricyclic Ligands and the Complexes with Halide Anions: F⁻, Cl⁻, Br⁻*. Journal of Computational Chemistry. **31**(4): p. 871-881.
15. Schwenke, D.W. and D.G. Truhlar, *Systematic study of basis set superposition errors in the calculated interaction energy of 2 HF molecules*. Journal of Chemical Physics, 1985. **82**(5): p. 2418-2426.
16. Kubicki, J.D., G.A. Blake, and S.E. Apitz, *Molecular orbital calculations for modeling acetate-aluminosilicate adsorption and dissolution reactions*. Geochimica Et Cosmochimica Acta, 1997. **61**(5): p. 1031-1046.

17. Dewar, M.J.S., et al., *AMI - A new general purpose quantum mechanical molecular model*. Journal of the American Chemical Society, 1985. **107**(13): p. 3902-3909.
18. Stewart, J.J.P., *Optimization of parameters for semiempirical methods. I. Method*. Journal of Computational Chemistry, 1989. **10**(2): p. 209-220.
19. Jensen, F., *Introduction to Computational chemistry*. 2007: Wiley.
20. Møller, C. and M.S. Plesset, *Note on an Approximation Treatment for Many-Electron Systems*. Physical Review, 1934. **46**(7): p. 618.

3 The Heteronaphthalene Potential Alphabet Letter Set

3.1 Designing a Heteronaphthalene Potential Alphabet

In order to try to gain some insight into nature's choice of nucleotides, a set of molecules (letters) will be constructed and studied. This set of molecules will be similar to nucleotides in that each molecule can be assigned a 4-bit numerical representation but in contrast to nucleotides (in which a combination of D/A pattern (3-bits) and size (1-bit) of each molecule are used to assign a binary pattern (see section 1.3)) molecules in the set to be proposed here will take all 4 data bits directly from the D/A pattern.

In conceiving an alternate alphabet we consider a system similar to DNA with which we are already familiar. We imagine that the alternate alphabet will have a backbone structure and replication method analogous to that of DNA. Individual molecular associations are thought of as occurring within a constraining environment equivalent to polymerase. The composition of the backbone or the polymerase equivalent environment will not be explored in this thesis. To do so, apart from the computational complexity involved and the resources required, would risk introducing artefacts as any choice made could distort the results determined for interactions between the molecules expressing information. Instead of modelling a specific backbone structure or polymerase analogous environment, molecular geometry constraints will be used as a proxy for a constraining environment. In this chapter we consider the construction of a 4 D/A position alphabet and how best to model it. The proposed set of molecules will act as an ideal set and will be designed in such a way that the molecules are kept uniform in structure. With this consistency in structure in mind it was decided that due to its rigid structure and shape and its ability to accommodate 4 hydrogen D/A sites, a naphthalene structure (Fig. 3.1) will be used as the basic template for each molecule. Each molecule contains an R group position (Fig. 3.1) at which a backbone structure would theoretically be attached if one were to be considered. In order to explore the entire set of 4 D/A position structures 16 ($2^4 - 2$ binary digits and 4 possible positions) molecules must to be constructed, one for each unique D/A pattern.

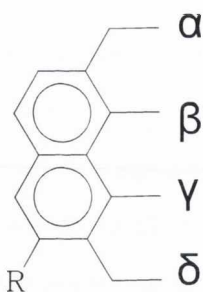


Figure 3.1 Basic naphthalene template. Positioning of R group indicated.

In constructing each molecule the following rules were used during design;

- A lone-pair (Lp) external to the ring is provided by a carbonyl group
- A H external to the ring is provided by NH_2
- Where possible a Lp in the ring is expressed by a nitrogen atom
- A H in the ring is expressed by NH

An example of construction can be seen for pattern 0100 (Fig. 3.2). The D/A pattern is first arranged using the design principles outlined above (D/A structure shown in the molecule on the left in Fig. 3.2), carbonyl groups are used to give terminal position lone-pairs (0), a nitrogen atom is used within the ring to give a 0 and an NH is used to give a 1 in a middle position. Once the correct D/A pattern is in place the rest of the molecular structure is completed with hetero atoms as required to satisfy valency (full structure shown on right Fig. 3.2).

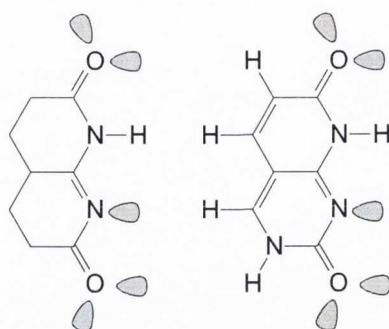


Figure 3.2 Pattern 0100 template and finished molecule.

All associations needed to complete the alphabet were constructed in the same way. The completed set of 16 molecules is termed heteronaphthalenes (Het) for convenience (Fig. 3.3). Each molecule is given a label from A-H: the presence of an * indicates a complementary letter. In this notation each letter has a complementary letter in which the D/A pattern is opposite in all positions. For example, F 0101 has a complement F* with a D/A pattern of 1010. Each of the 8 letters labelled A-H is linked to a complementary letter of opposite D/A pattern indicated using the same label with the presence of an *. A complementary pair is indicated using both letters involved in the molecular pair, for example, Het[FF*] notates the pair formed between F 0101 and its complementary letter F* 1010.

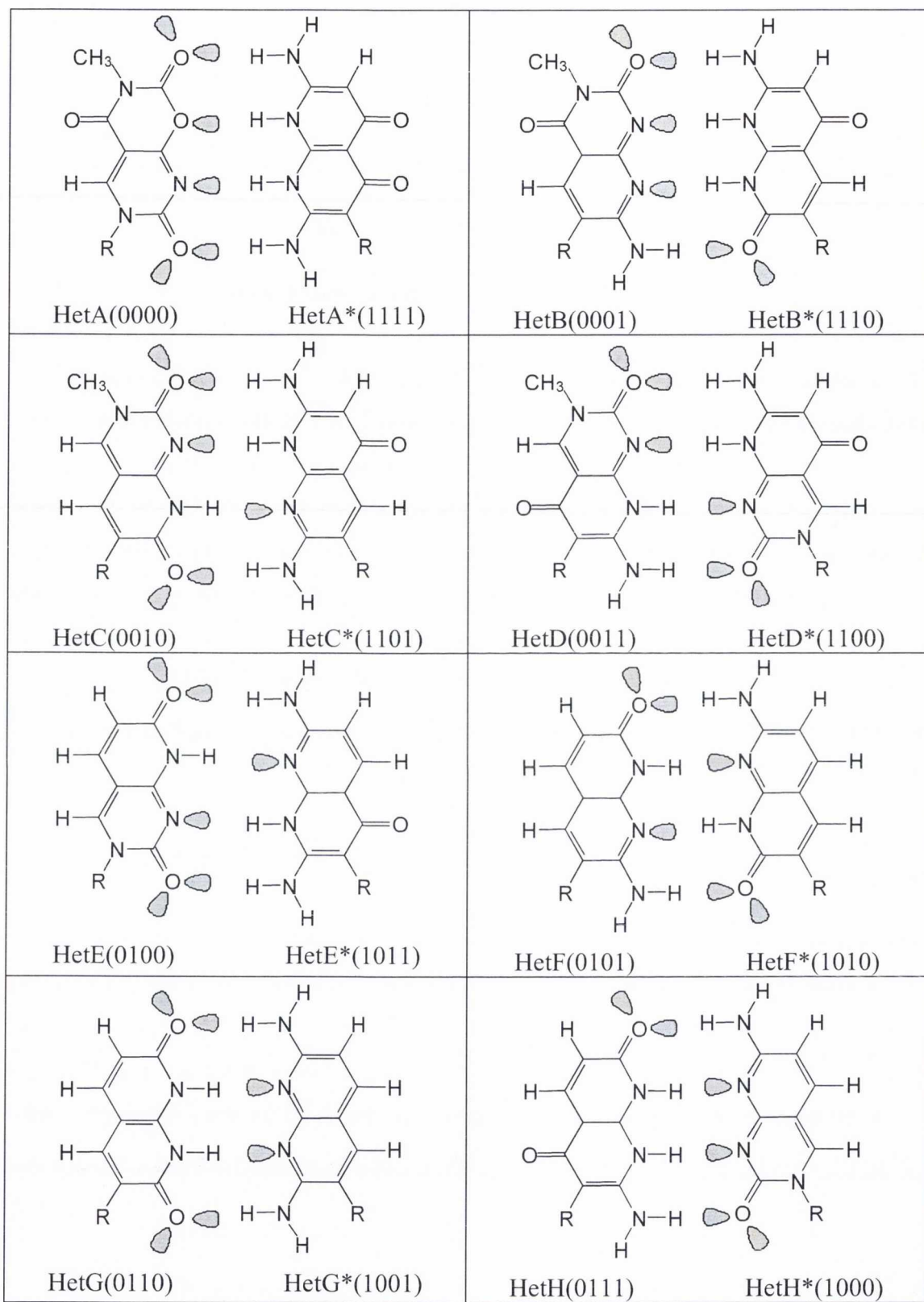


Figure 3.3 The 16 possible 4-bit hydrogen/lone pair patterns expressed in heteronaphthalenes

Each molecule has an R group (as depicted in Fig. 3.3) representing the point at which the molecule would theoretically be attached to the backbone structure (imagined to be analogous to the phosphate-deoxyribose backbone in DNA). In this study of the Het potential alphabet letter set a hydrogen atom is used as the R group. As the Het alphabet will act here as an ideal alphabet to be compared to a more realistic one later in this thesis, using a hydrogen atom in the R group position makes uniformity at the point of hypothetical attachment easily achievable.

3.2 Exploring a potential heteronaphthalene alphabet

Investigating whether the proposed Het alphabet meets the necessary conditions for viability (see section 1.4) requires the study of all possible pairings, those that are complementary and those described as mismatching. A mismatch can be either lone pair - lone pair (Lp-Lp) or hydrogen - hydrogen (H-H). Each Het molecule has four distinct hydrogen bonding positions (labelled α , β , γ and δ (see section 3.1)), in each of which a match or mismatch can be present, 5 types of interaction need to be considered;

- Zero mismatches (complementary associations) $\bar{\partial}=0$
- Mismatches in one positions $\bar{\partial}=1$
- Mismatches in two positions $\bar{\partial}=2$
- Mismatches in three positions $\bar{\partial}=3$
- Mismatches in four positions $\bar{\partial}=4$

In molecular recognition the ability of a molecule to discriminate between complementary and non-complementary letters is linked to the geometric and spatial freedom which the molecule has. Non-complementary associations which in the standard Watson-Crick arrangement are repulsive may find a stable binding conformation if left completely free. One example of this is ‘wobble’ which occurs by shifting the position of one molecule relative to the other, and in doing so can create one or more new matches (Fig. 3.4).

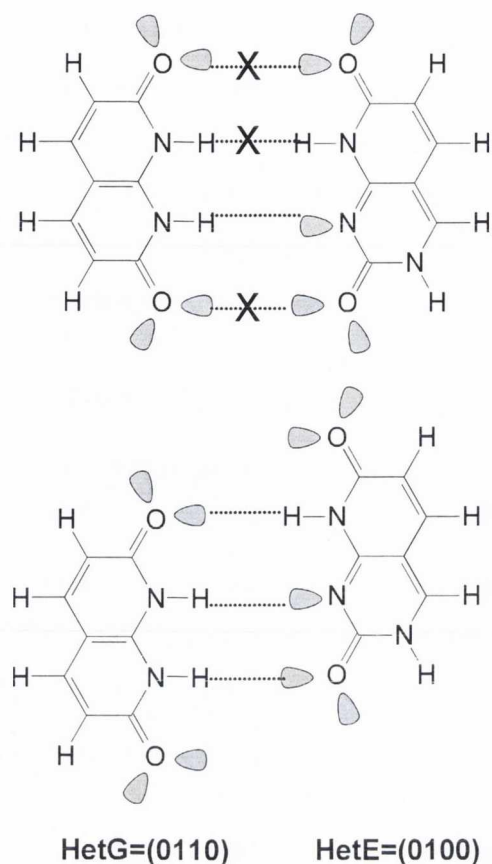


Figure 3.4 Example of wobble, moving down in just one position causes three mismatches to become 3 matches.

To prevent wobble as well as other ways in which molecules could move to avoid mismatch repulsion a set of molecular geometry restrictions will be designed. The restrictions will be designed such that complementary pairs fit comfortably within them and non-complementary pairs do not. In this way the constraints used will work as a selecting agent discouraging non-appropriate associations by placing them into a restricted space thus causing them to encounter strong steric repulsions. The greater the repulsion experienced between non-complements the less likely it is for a mismatched pair to occur.

For the extended nucleotide alphabet it has been seen that not all letters form equally viable sub groups of letters. It is predicted that the same will be true for the proposed Het alphabet. One of the aims in the study of this Het potential alphabet letter set will therefore be to explore if any subsets of letters exist that could perhaps form a viable alphabet.

3.2.1 Construction of molecular geometry constraints

For the proposed potential Het alphabet, geometric parameters which correspond to a space into which complements fit comfortably need to be determined. In order to determine appropriate restriction values, a study was undertaken exploring the relaxed geometry of the 8 complementary pairs. Geometry optimizations were performed using Hartree-Fock (HF), 6-31G* basis set with Gaussian 03W. The only geometric restriction in place is the C_s point group, thus keeping both molecules confined to the molecular plane. To do otherwise would raise further questions regarding how much out of plane freedom molecules should be given which could not easily be answered and could introduce further artefacts that may possibly, obscuring the results of this ideal Het alphabet study.

The results shown below indicate that in all cases except for pair Het[AA*] 0000-1111 (which shows a degree of wobble) a Watson-Crick alignment is maintained. The bond distances are measured from heavy atom to heavy atom (N---N or N---O) and angles can be seen in Table 3.1 and Table 3.2 respectively. The bond lengths seen across the pairs are largely consistent ranging from 2.86 Å – 3.20 Å (3.56 Å Het[AA*]) (for further discussion of bond lengths see appendix A1). In general the middle two positions sit further apart compared to the more flexible terminal positions. It is noted that the hydrogen bond angles presented here have been measured using the measured angle function integrated into the GaussView programme. This function measures how far an angle deviates from 180° . This means that an angle of 175° and 185° would both be given the same value (Fig. 3.5).

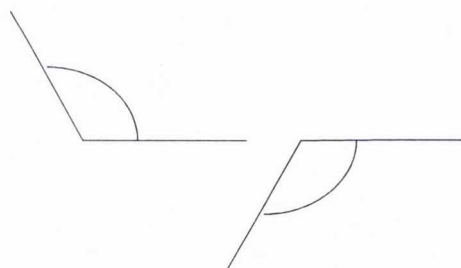


Figure 3.5 Angles less than and greater than 180° when equal in magnitude are given the same value using GaussView. Angles here are not shown to scale.

Het[AA*] [0000-1111] was excluded from the calculation of average bond lengths and angles as it deviates from all the other pairs. It should be noted that Het[AA*] deviates in structure from the other Het pairs, it contains an oxygen within the ring structure giving it

three N-H---O bonds and only one N-H---H. The other seven pairs contain the two bond types in an equal mix (2 N-H---N, 2 N-H---O). It is unlikely that the removal of a single outlier will have a significant effect in determining suitable geometric conditions which will be used to mimic a constraining environment. Including outliers that would allow pairs greater freedom of movement could in fact increase the possibility of a mismatching association fitting comfortably into the pocket.

Table 3.1 Geometry of complementary pairs optimised with HF 6-31G(d) basis set. The four hydrogen bond distances in α β γ δ positions are measured for each pair.

HF 6-31G* Free	α (Å)	β (Å)	γ (Å)	δ (Å)	Average $\beta\gamma$ (Å)
Het[AA*] 0000-1111	3.25	3.56	3.43	3.14	3.49
Het[BB*] 0001-1110	3.01	3.13	3.10	2.86	3.11
Het[CC*] 0010-0010	3.05	3.20	3.18	3.03	3.19
Het[DD*] 0011-1100	2.95	3.09	3.08	2.97	3.09
Het[EE*] 0100-1011	3.04	3.15	3.18	3.03	3.17
Het[FF*] 0101-1010	2.97	3.20	3.20	2.97	3.20
Het[GG*] 0110-1001	2.93	3.11	3.11	2.93	3.11
Het[HH*] 0111-1000	2.86	3.10	3.12	3.03	3.11
Average over all positions	3.01	3.19	3.17	3.00	3.18
Average excluding Het[AA*]	2.97	3.14	3.14	2.97	3.14

Table 3.2 Geometry of complementary pairs optimised with HF 6-31G(d) basis set. The four hydrogen bond angles in α β γ δ positions are measured for each pair.

HF 6-31G* Free	α (degrees)	β (degrees)	γ (degrees)	δ (degrees)	Average $\beta\gamma$ (degrees)
Het[AA*] 0000-1111	167.725	168.192	175.621	178.391	171.907
Het[BB*] 0001-1110	179.872	177.432	174.017	178.690	175.725
Het[CC*] 0010-0010	176.150	178.381	178.751	178.506	178.566
Het[DD*] 0011-1100	179.416	175.980	175.982	179.446	175.981
Het[EE*] 0100-1011	175.640	178.151	178.046	179.301	178.099
Het[FF*] 0101-1010	178.854	175.806	175.806	178.856	175.806
Het[GG*] 0110-1001	179.788	178.303	178.294	179.776	178.298
Het[HH*] 0111-1000	178.689	173.662	177.123	179.913	175.393
Average over all positions	177.017	175.738	176.705	179.110	176.222
Average $\beta\gamma$ excluding Het[AA*]	178.344	176.816	176.860	179.212	176.838

In constructing the set of constraints it was decided to constrain only the middle two hydrogen bonding positions to allow the pair as much movement as possible within the pocket. The average bond length over β and γ excluding Het[AA*] was determined to be 3.14Å. Constraining bond distances alone is insufficient to prevent molecules from sliding apart to minimise repulsions. To address this the bond angles of the middle two positions

are also frozen. It was decided to freeze the internal hydrogen bond angles at 180° even though the average $\beta\gamma$ angles differs from this. Locking angles is not as simple as locking a bond distance, it could validly be done from either side of 180° as depicted in Fig. 3.5, thus increasing the chance of inconsistency throughout a set of results. Choosing 180° removes any chance of introducing an inconsistency of this type. This combination of bond and angle restrictions will form the standard geometry constraints (STRD) and are used to represent a constraining environment.

Now that suitable molecular geometry restrictions have been determined they can be used as a selection tool in the exploration of the complete set of Het letters. The STRD molecular geometry constraints will be use in the study of all pairs complementary or not. To summarize the geometric STRD constraints are defined as detailed in Table 3.3.

Table 3.3 STRD geometry restrictions for Het associations

Heavy atom distance of 3.14 Å in the middle two positions
Bond angles locked to 180° in the middle two positions
The C_s point group

3.3 Complementary Heteronaphthalene associations-Results

In order to assess the effect of the STRD constraints the 8 complementary Het associations were explored using the proposed conditions. The results can then be directly compared to the free (apart from the C_s point group) associations. The total interaction energy (TIE - the difference between monomers and dimer when the monomers have geometry independent of the dimer) was calculated for each of the pairs using HF 6-31G*: BSSE has not been removed at this stage (Table 3.4)(Fig. 3.6). In the study of this idealised Het alphabet it is the overall results that are important rather than any individual interaction energy. On comparing the TIEs for free and STRD geometry complementary pairs it is noted that with the geometry restrictions in place all of the energies have become more repulsive but the overall trend in values remains unchanged. With the exception of Het[AA*] only a small difference is noted in the energies calculated with and without STRD conditions in place, which indicates that the geometric pocket comfortably fits complements.

Figure 3.4 Het TIEs for free and standard (STRD) conditions

Pair	TIE Free (kJ/mol)	TIE STRD (kJ/mol)	Difference (kJ/mol)
Het[AA*] 0000-1111	-73.522	-57.718	15.804
Het[BB*] 0001-1110	-160.883	-159.61	1.273
Het[CC*] 0010-0010	-88.961	-88.338	0.623
Het[DD*] 0011-1100	-147.449	-146.232	1.217
Het[EE*] 0100-1011	-92.783	-92.378	0.405
Het[FF*] 0101-1010	-88.165	-86.551	1.614
Het[GG*] 0110-1001	-124.781	-124.236	0.545
Het[HH*] 0111-1000	-159.981	-158.645	1.336

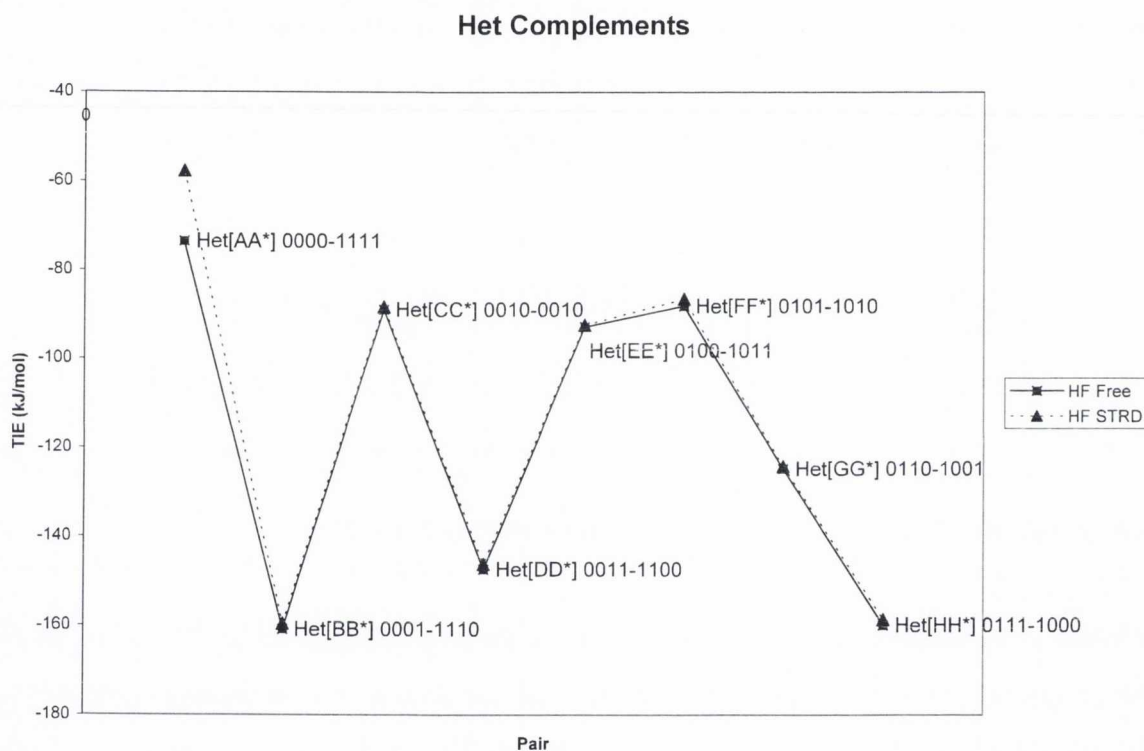


Figure 3.6 Plot of Het TIEs for free and standard (STRD) conditions.

A range of TIE values is seen for the complementary pairs (-58 to -159 kJ/mol).

This range of values indicates that although each of the associations is similar in structure and has an equal number of hydrogen bonds, something else may be systematically affecting their relative stabilities, if this were not the case a plot analogous to that shown in Fig. 3.7.

The variation seen in TIE values could be due to the non equivalence of the hydrogen bonds that exist between adjacent atoms (commonly referred to as secondary interactions).

Before proceeding to consider mismatching associations we first attempt to rationalize the considerable variation present in the calculated energies using the secondary interactions [1].

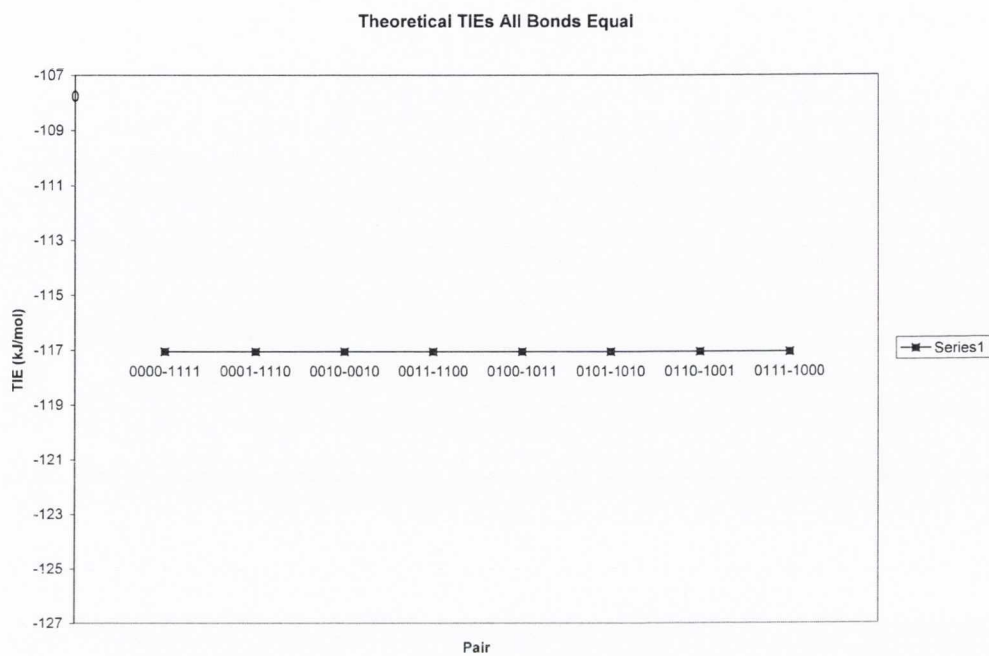


Figure 3.7 Theoretical plot showing the overall trend if all hydrogen bonds were equal
In the plot the average the TIE over all 8 free associations has been used.

1. Jorgensen, W.L. and J. Pranata, *Importance of Secondary Interactions in Triply Hydrogen-Bonded Complexes - Guanine-Cytosine Vs Uracil-2,6-Diaminopyridine*. Journal of the American Chemical Society, 1990. **112**(5): p. 2008-2010.

4 Secondary Interactions

4.1 Introduction

In our genetic alphabet nucleotide base pair C:G is seen to be more binding than A:T (Fig. 4.1) [1,2].

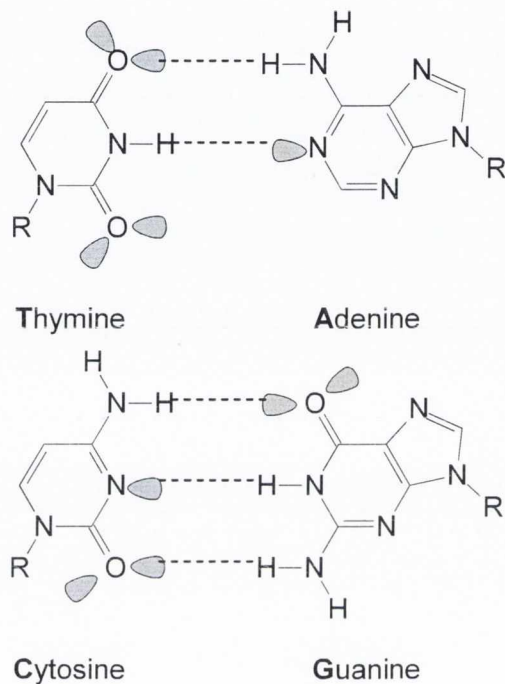


Fig. 4.1 Base pairs in DNA.

At first glance the relative binding strength of C:G compared to A:T could be presumed to be due to the fact that it has three hydrogen bonds whilst A:T has only two. Jorgensen et al. [3] explored the binding strengths of the two nucleotide pairs by comparing C:G to U:DAP (Diaminopyridine) (Fig. 4.2), both of which have three hydrogen bonds, and showed that the answer is not simply a consequence of the number of hydrogen bonds present. They saw that even when both pairs had an equal number of hydrogen bonds, C:G still had a significantly higher binding energy (C:G 92.47 kJ/mol, U:DAP 47.70 kJ/mol).

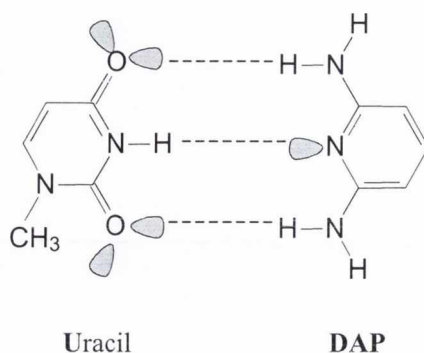


Figure 4.2 U-DAP

The increased relative binding strength of C:G was explained by Jorgensen et al. when the secondary interactions (SI) of the hydrogen bonds were taken into account. SI arise from interactions that occur adjacent to primary hydrogen bonds. Nucleotide base pair C:G, as shown below (Fig 4.3) has three primary hydrogen bonds and also four secondary interactions. SI can be stabilizing or destabilizing. In the case of C:G two of the SI where the hydrogen bond dipoles are aligned are attractive (stabilizing hydrogen-lone pair), represented by a block line and two when the dipoles are opposite are repulsive (destabilizing hydrogen-hydrogen or lone pair-lone pair) represented with a dashed line, in total this leaves zero net secondary interactions.

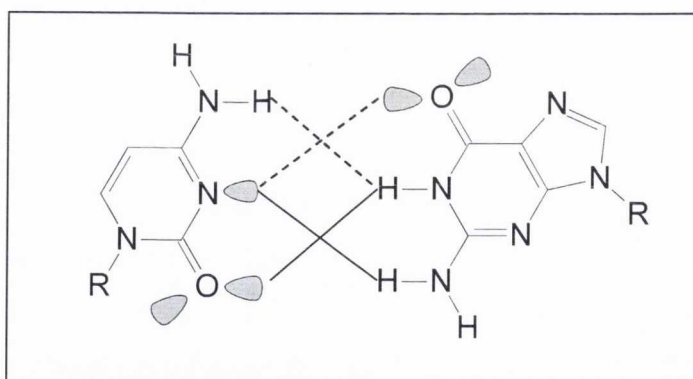


Figure 4.3 SI shown for nucleotide base pair C:G.

2 attractive and 2 repulsive: Net=0

SI can produce a net stabilizing or destabilizing value and do not always cancel out. In U:DAP (Fig. 4.4) all four secondary interactions are repulsive, leading to U:DAP having a lesser binding energy when compared to C:G. This nucleotide example illustrates how important the contribution from SI can be to the association energy of a molecular dimer.

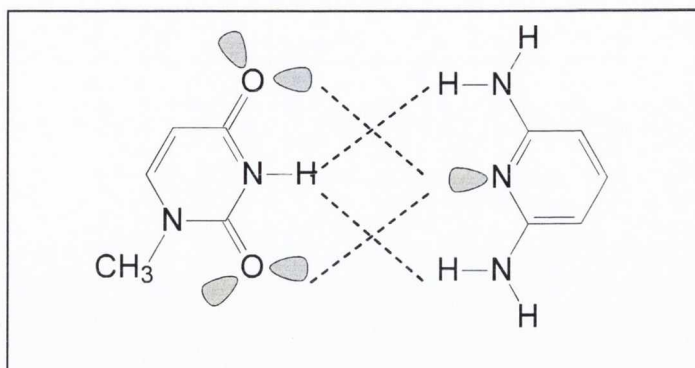


Figure 4.4 SI for U:DAP.

Net= -4

Post the original work of Jorgensen and Pranata values have been fitted in the literature for primary hydrogen bonds and secondary interactions. Pranata et al. [4] analysed the interactions of C:G and D:UAP and determined values for primary (-31.38 kJ/mol) and secondary (+/- 10.46 kJ/mol) hydrogen interactions. Sartorius and Schneider [5] used a range of D/A complexes with varying numbers of primary hydrogen bonds and determined a value for primary (-7.9 kJ/mol in solution) and secondary (+/- 2.9 kJ/mol) interactions. They noted a good agreement between their approach of taking only one value for primary interactions and one for secondary and experimentally determined results.

Although the Jorgensen et al. SI model is often referred to and used in the literature [6-10], it has not universally been seen to explain data trends, leading some authors to suggest it be applied with caution [11, 12].

Now that we have seen that it is not just primary hydrogen bonds that determine the energy of a molecular pair, further exploration can be undertaken into the spread of energy values seen for the proposed Het alphabet.

4.2 Heteronaphthalene secondary interactions-Results

In section 3.3 the TIE for each of the 8 complementary Het associations were determined and although each pair contains an equal number of hydrogen bonds a spread of values was seen. In addition to the four primary hydrogen bonds each association also has six SI (Fig. 4.5)(Table 4.1).

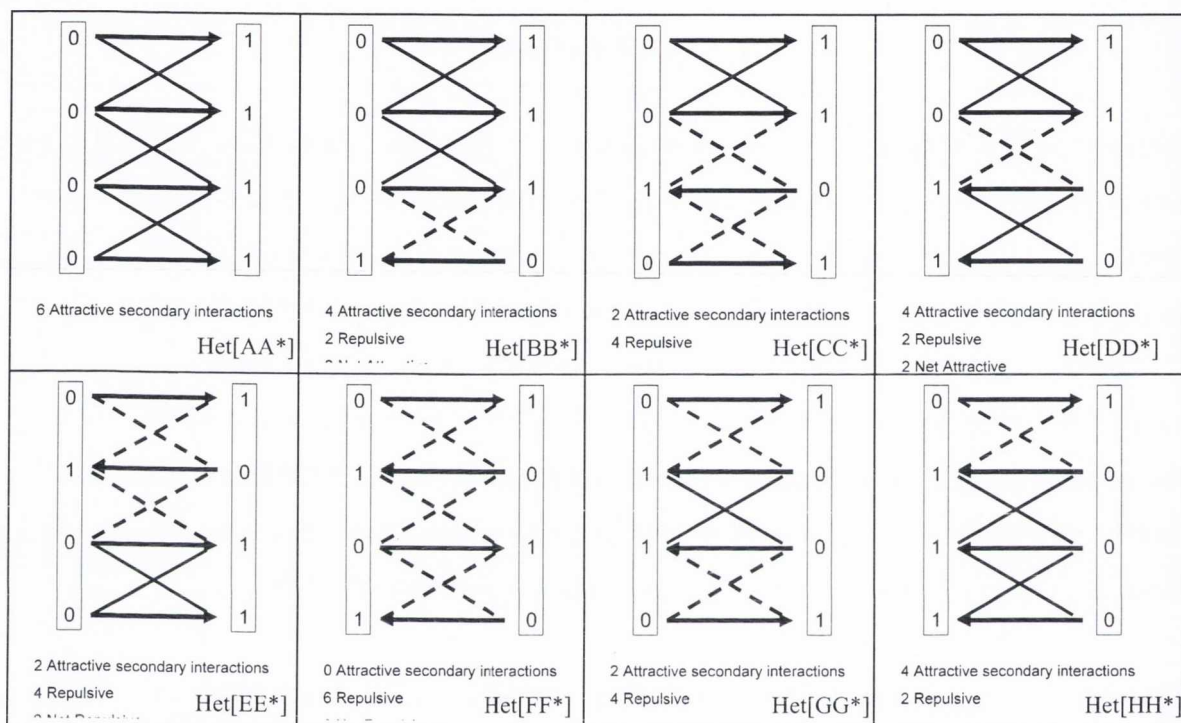


Figure 4.5 Primary (indicated by the presence of a solid arrow) and secondary interactions (attractive shown using a solid line, repulsive using a dotted line) for all 8 Het complementary pairs.

Table 4.1 Ordering of pairs based on SI present

Pair	Net SI
Het[AA*] 0000-1111	6
Het[BB*] 0001-1110	2
Het[HH*] 0111-1000	2
Het[DD*] 0011-1100	2
Het[CC*] 0010-0010	-2
Het[EE*] 0100-1011	-2
Het[GG*] 0110-1001	-2
Het[FF*] 0101-1010	-6

A plot can be made of the net number of SI per pair (Fig. 4.6). This SI plot closely matches the overall shape of that seen for the TIEs of the Het complementary associations (Fig. 4.7). The only major departure from the predicted trend is Het[AA*] 0000-1111 which is

predicted to be the most binding pair. Its deviation may be somewhat due to HetA being the only letter with an oxygen within its rigid ring structure giving it three N-H---O bonds.

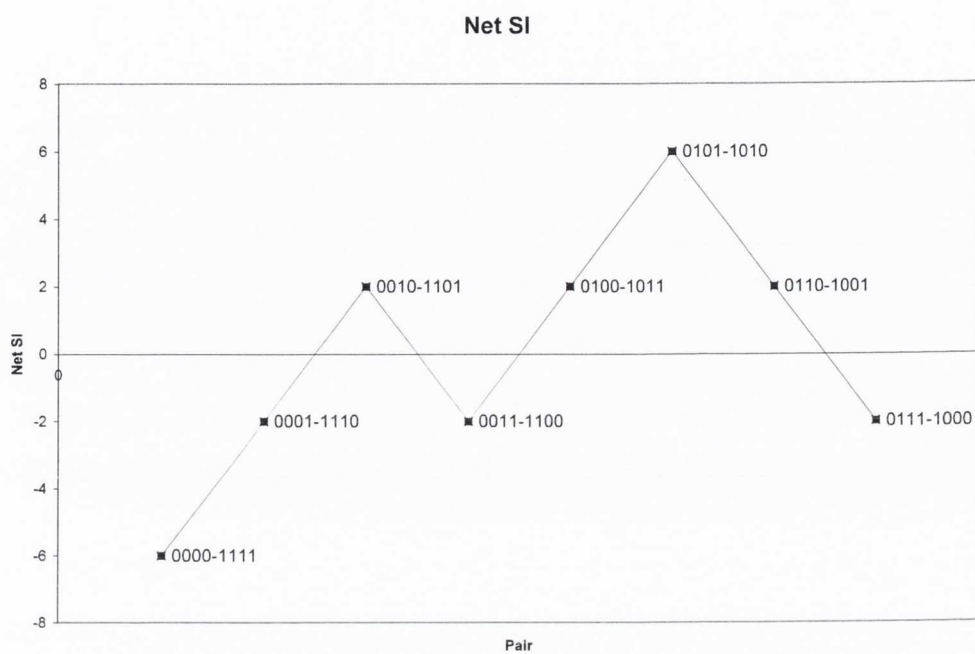


Figure 4.6 Plot of SI per pair.

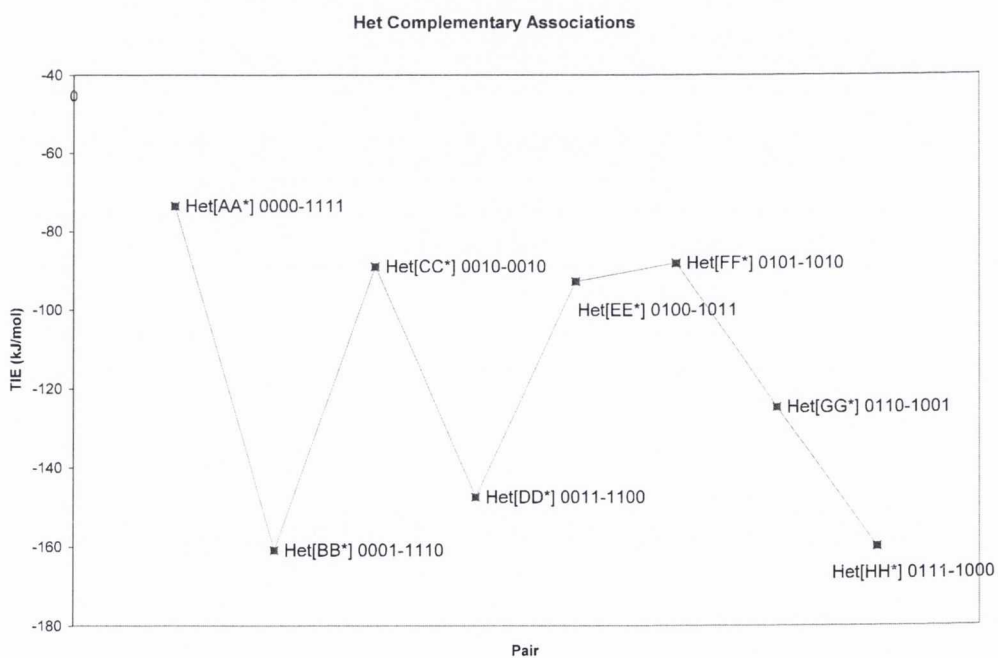


Figure 4.6 TIEs Het Free geometry

This similarity in trends suggests that SI could in fact in this case account for the variation in interaction energies seen for complementary associations. In order to explore this further and to determine if secondary interactions can be used to predict TIEs for hydrogen bonding arrays (similar to the work of Pranata et al. [4], Sartorius and Schneider [5]), average values for primary and secondary interactions can be determined using the complete set of results and then used to predict a new set of results (see appendix A4). The average primary hydrogen bond was calculated to be -31.786 kJ/mol and the average secondary ± 10.488 kJ/mol. These predicted energies for the Het associations and those from literature can be plotted and compared to HF calculated energies (Fig. 4.8).

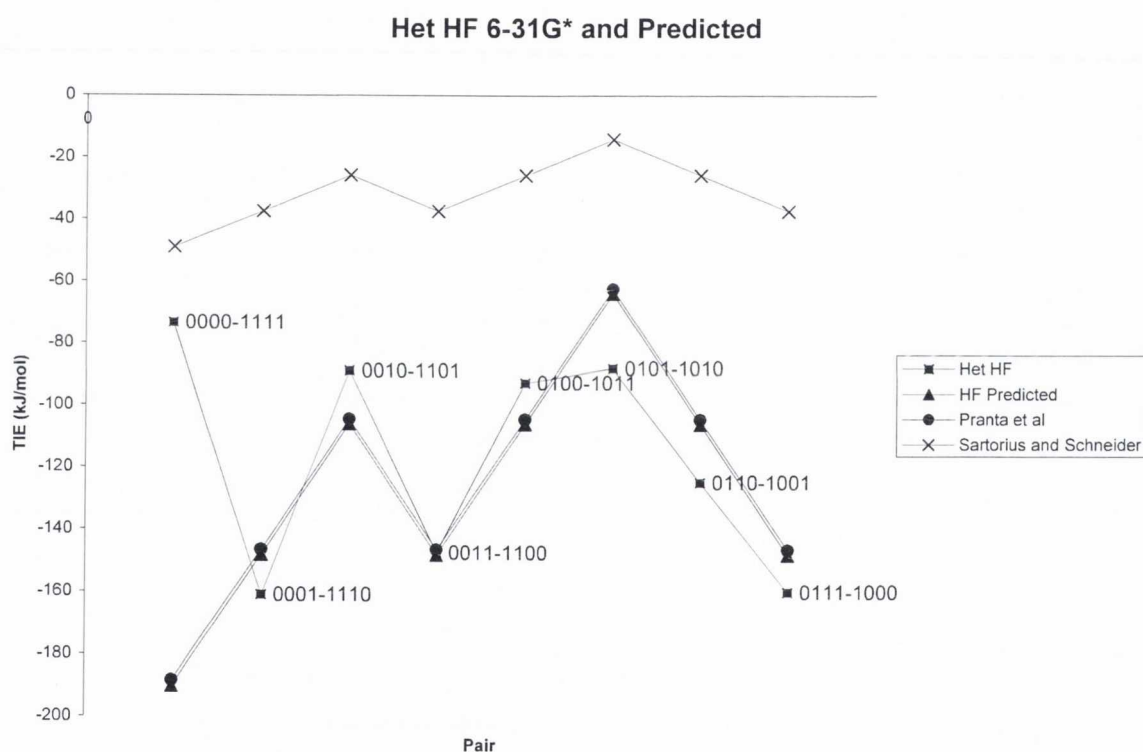


Figure 4.8 TIE for each Het pair, predicted and HF calculated.

All of the predicted plots (literature and predicted directly from the Het results) show the same pattern of results, although the pattern is shifted depending on which prediction is used. This shift in results depending on which prediction is used is to be expected as the literature values are based on different systems and use different methods or derivation. The Pranata et al. [4] prediction (based on nucleotides) very closely matches that determined for the Het potential alphabet. In all cases deviation from the secondary

interaction trend prediction most notably at Het[AA*] 0000-1111. Other deviations can be noted between pairs with equal net secondary interactions but different D/A patterns, for example, Het[BB*] and Het [DD*].

4.2.1 A Heteronaphthalene based fit for secondary interactions

Deviations from the prediction SI behaviour are noted in pairs with the same number of net SI but different D/A patterns. If the patterns with the same number of net secondary interactions are examined, two groups of associations emerge;

{Het[BB*],Het[DD*],Het[HH*]} (Fig. 4.8)(Table 4.2) where each pair has 2 net secondary interactions and {Het[CC*],Het[EE*],Het[GG*]} (Fig. 4.9)(Table 4.3) in with each pair has -2 net secondary interactions Examining the secondary interactions for the first group in which each pair has 2 net interactions, two of the patterns 0001-1110 and 0111-1000 show essentially the same arrangement of secondary interactions. Both of these patterns have a block of 4 positive interactions followed (or preceded) by a block of 2 negative. The third pattern 0011-1100 shows a different arrangement, a block of two positive interactions followed by a block of 2 negative and ending with another block of 2 negative.

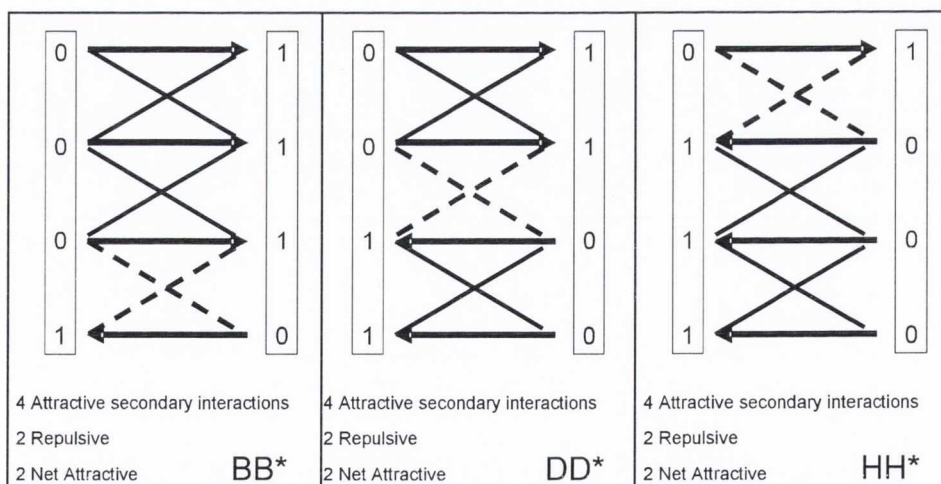


Figure 4.8 SI [Het[BB*],Het[DD*],Het[HH*]]

Table 4.2 TIE [Het[BB*],Het[DD*],Het[HH*]]

Pair	TIE HF 6-31G*
Het[BB*] 0001-1110	-160.883
Het[DD*] 0011-1100	-147.449
Het[HH*] 0111-1000	-159.981

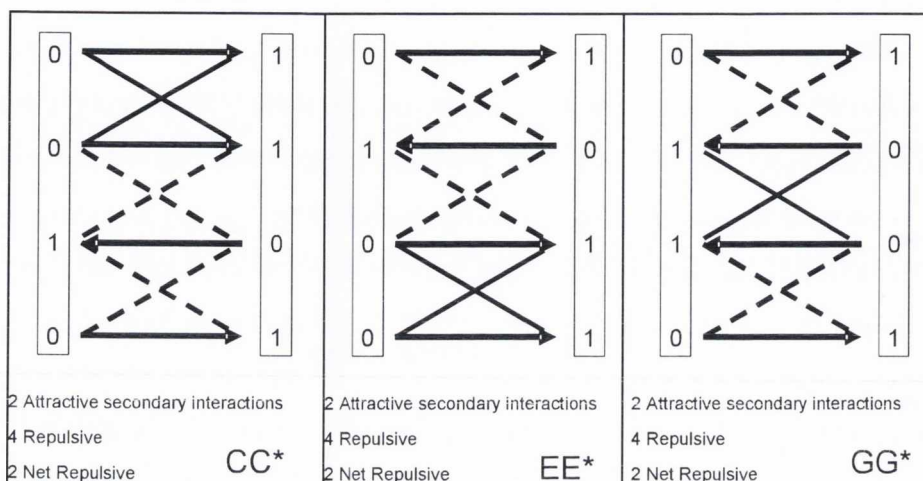


Figure 4.9 SI [Het[CC*],Het[EE*],Het[GG*]]

Table 4.3 TIE [Het[CC*],Het[EE*],Het[GG*]]

Pair	TIE HF 6-31G*
Het[CC*] 0010-0010	-88.961
Het[EE*] 0100-1011	-92.783
Het[GG*] 0110-1001	-124.781

Looking at the TIEs calculated for pairs in the two groups, Het[DD*] shows a less binding interaction energy compared to the other pairs in its group. The reverse is seen for Het[GG*] which shows a stronger interaction energy compared to Het[CC*] and Het[EE*]. In each set the largest difference in interaction energy is noted in the pattern with the “sandwiched” positive or negative block occurring in the middle position. This finding suggests that not all secondary interactions are in fact equal, their strength could be dependant on positioning. The possibility of “sandwiching” does not exist for pairs with three or primary hydrogen bonds. In the case of three positions only three ways of arranging secondary interaction exist (Fig. 4.9).

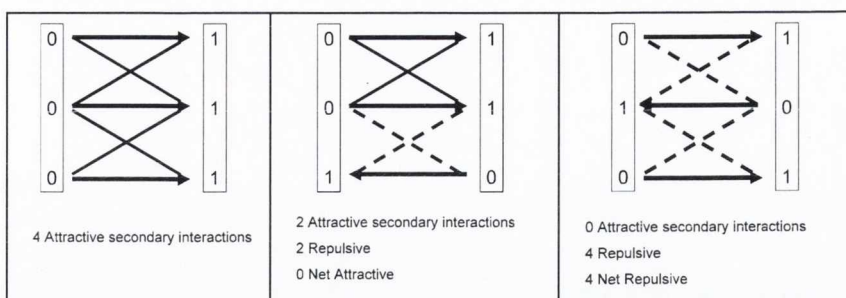


Figure 4.9 All possible hydrogen bonding SI for triply hydrogen bond arrays.

The basic SI model presumes that all primary hydrogen bonds are equivalent in value and also all SI (stabilizing or destabilizing) also have only one value (see section 4.1). This is not the case as some hydrogen bond types are stronger than others [13][14]. Variation in hydrogen bond strengths was taken into account by Quinn et al. during a study of C-H---O interactions [15], where they devised an equation giving different bond types weighted values. This weighted model would have little effect in the case of the Het potential alphabet set of letters as 7 out of the 8 pairs contain 2 N-H---N bonds and 2 N-H---O bonds, thus weighting the bonds in these bases would have no effect.

A new way of looking at secondary interactions in the Het alphabet can be devised in which the positioning of the interactions present will also be taken into account. It is hoped that in doing this the SI model can be made a better fit for the Het set. Using this new analysis an equation can be constructed and used to summarise the primary and secondary interactions for a given pattern. An example of this can be seen below for pattern 0001-1110 (Fig. 4.10). In this pattern there are four primary hydrogen bonds each denoted by P, three sets of secondary hydrogen bonds, each set is assigned either as aligned A (cross terms are positive) or not aligned NA (if the cross interactions are negative).

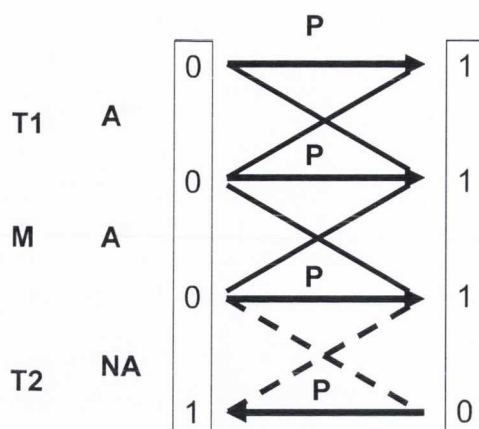


Figure 4.10 D/A pattern 0001-1110

In the set of Het patterns being considered, each pair has 4 primary (4P) hydrogen bonds and 2(P-1) secondary interactions. The three secondary bonding blocks are labelled according to positions: T1 represents the top terminal, M middle and T2 the bottom terminal. Using the above notation the number of net interactions (NI) (primary and secondary) for the pair 0001-1110 can be described as (Eqn. 4.1);

$$NI = 4P + 2(A(T1) + A(M) + NA(T2)) \quad \text{Equation 4.1}$$

In this first instance we will assume that all secondary interactions are equal, if this is the case $A = -NA$ and $T1 = M = T2$. For the sample pattern 0001-1110 (Eqn.4.2),

$$NI = 4 + 2(1 + 1 + (-1)) \quad \text{Equation 4.2}$$

As each of the complementary Het associations has four primary hydrogen bonds these can be omitted giving the pattern 0001-1110 and $NI = 2$.

For the Het pairs the middle interaction positions are the most rigid (with and without STRD geometry restrictions in place) the calculation of NI can be modified in order to reflect this. A series of different weighting factors was applied to the interaction in the M (middle) interaction block. The optimum weighting value was found (through systematic trial) to be 1.5 for each of the middle interactions (Eqn.4.3).

$$NI (M=1.5) = 4 + (T1) + (-1.5)(M) + (T2) \quad \text{Equation 4.3}$$

In the pattern 0011-1100 for example (primary hydrogen bonds are again neglected as these are uniform across the alphabet) (Eqn.4.4);

$$NI (M=1.5) = 4 + 2(1(T1) + (-1.5)(M)+1(T2))=5 \quad \text{Equation 4.4}$$

This new weighting factor can be applied to all pairs in the alphabet and new plots based on the number of SI made ($M = 1.5$) and then compared to the original ($M=1$) (Fig. 4.11, Fig. 4.11a)(Fig. 4.12, Fig. 4.12a). In this plot primary interactions have been neglected as they are uniform across the alphabet. Het[AA*] 0000-1111 has not been included in the linear fitting procedure due to its anomalous behaviour: its inclusion could mask key data relating to the remaining data point.

With the weighting factor in place a large improvement can be seen in the linearity of the plot. It seems for the Het set of molecules that the order of the complementary pairs can be rationalized (except in the case of Het[AA*] which is anomalous) in terms of secondary interactions.

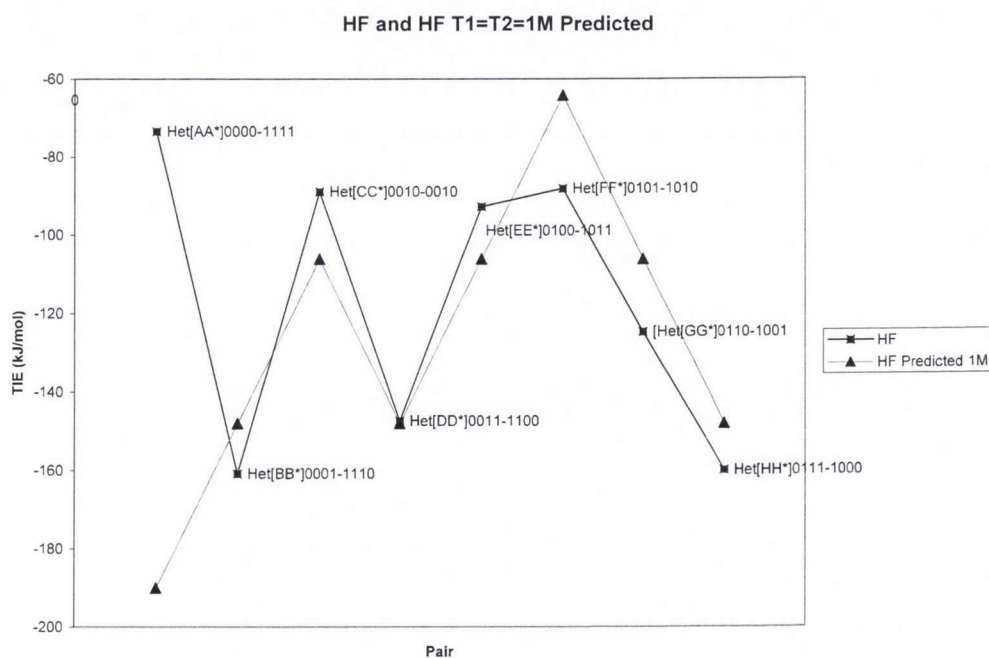


Figure 4.11 Het and Het predicted T1=T2=1M

Het Complements T1=T2=1M

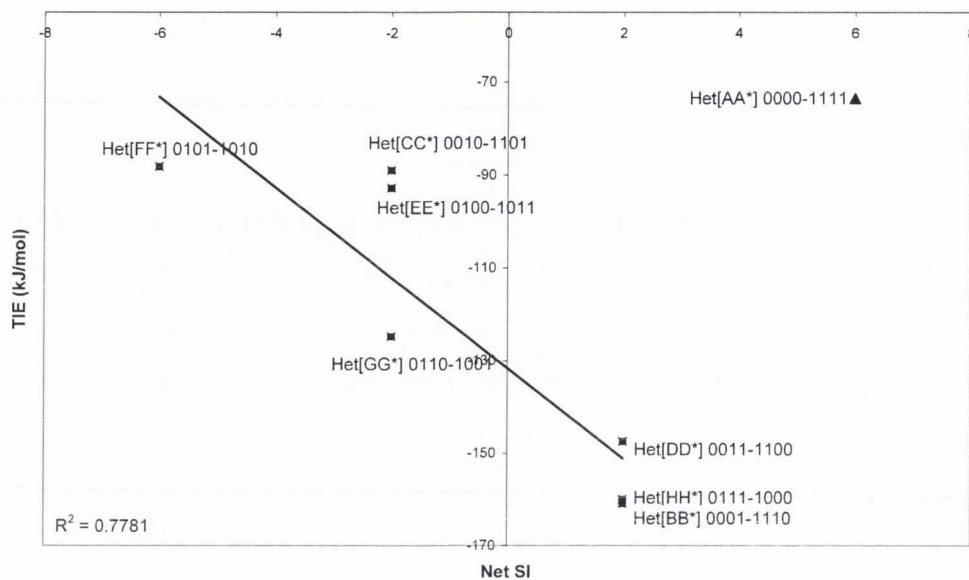


Figure 4.11a TIE Vs Net SI

HF and HF T1=T2=1.5M Predicted

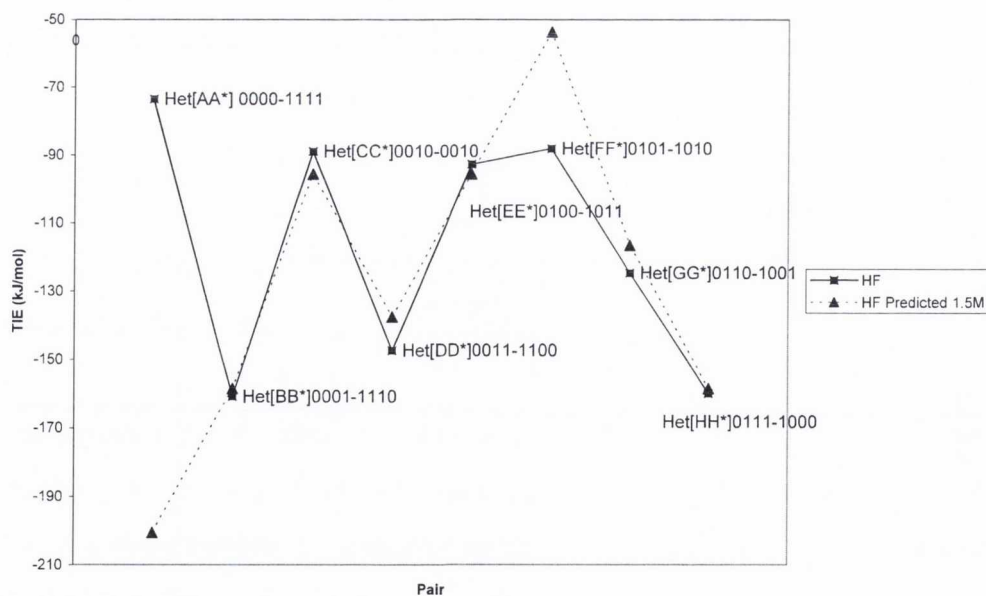


Figure 4.12 Het and Het predicted T1=T2=1.5M

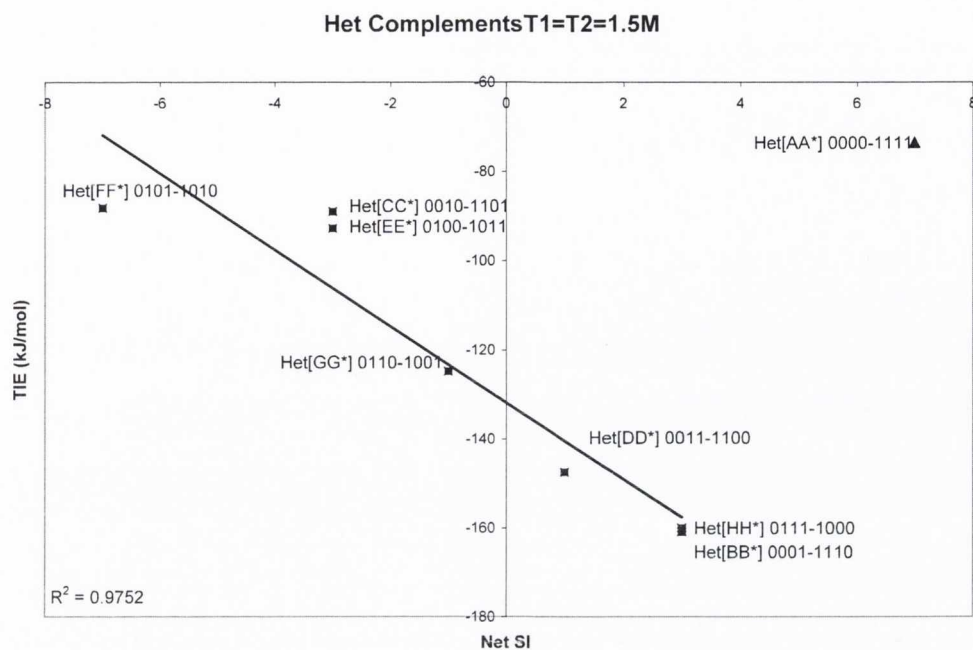


Figure 4.12a TIE Vs NI-P. Het 6-31G* Free T1=T2=1.5M

Overall the results indicate that using the number of secondary interactions as a rough guide can be loosely used to predict which hydrogen bonding array in a series (all things ideally being close to uniform throughout the series) will be the more binding. In highly rigid molecules such as the Het potential alphabet being considered here the predicted order of pairs with an equal number of SI can be improved by adding a weighting factor to allow for “sandwiching” of the interactions in the middle two positions. This increase in strength evident in the secondary interactions in the middle two secondary positions could be a consequence of the rigid structure of the Het molecules used and may not be as evident in a linear alphabet with greater freedom. This will be investigated later in this thesis for a potential alphabet letter set devised where possible from literature and also in brief for a linear chain-like D/A alphabet. Different calculation methods will also be considered.

In the next chapter we will systematically consider non-complementary Het associations, from pairs that mismatch in one position to those that mismatch in all four.

1. Kurita, N., V.I. Danilov, and V.M. Anisimov, *The structure of Watson-Crick DNA base pairs obtained by MP2 optimization*. Chemical Physics Letters, 2005. **404**(1-3): p. 164-170.
2. Herbert, H.E., et al., *Hydrogen-bonding interactions in peptide nucleic acid and deoxyribonucleic acid: A comparative study*. Journal of Physical Chemistry B, 2006. **110**(7): p. 3336-3343.
3. Jorgensen, W.L. and J. Pranata, *Importance of Secondary Interactions in Triply Hydrogen-Bonded Complexes - Guanine-Cytosine Vs Uracil-2,6-Diaminopyridine*. Journal of the American Chemical Society, 1990. **112**(5): p. 2008-2010.
4. Pranata, J., S.G. Wierschke, and W.L. Jorgensen, *Opls Potential Functions for Nucleotide Bases - Relative Association Constants of Hydrogen-Bonded Base-Pairs in Chloroform*. Journal of the American Chemical Society, 1991. **113**(8): p. 2810-2819.
5. Sartorius, J. and H.J. Schneider, *A general scheme based on empirical increments for the prediction of hydrogen-bond associations of nucleobases and of synthetic host-guest complexes*. Chemistry-a European Journal, 1996. **2**(11): p. 1446-1452.
6. Leach, A.R. and P.A. Kollman, *Theoretical investigations of novel nucleic-acid bases*. Journal of the American Chemical Society, 1992. **114**(10): p. 3675-3683.
7. Murray, T.J. and S.C. Zimmerman, *New Triply Hydrogen-Bonded Complexes with Highly Variable Stabilities*. Journal of the American Chemical Society, 1992. **114**(10): p. 4010-4011.
8. Beijer, F.H., et al., *Strong dimerization of ureidopyrimidones via quadruple hydrogen bonding*. Journal of the American Chemical Society, 1998. **120**(27): p. 6761-6769.
9. Sherrington, D.C. and K.A. Taskinen, *Self-assembly in synthetic macromolecular systems via multiple hydrogen bonding interactions*. Chemical Society Reviews, 2001. **30**(2): p. 83-93.
10. Blight, B.A., et al., *An AAAA-DDDD quadruple hydrogen-bond array*. Nature Chemistry, 2011. **3**(3): p. 244-248
11. Popelier, P.L.A. and L. Joubert, *The elusive atomic rationale for DNA base pair stability*. Journal of the American Chemical Society, 2002. **124**(29): p. 8725-8729.
12. Lukin, O. and J. Leszczynski, *Rationalizing the strength of hydrogen-bonded complexes. Ab initio HF and DFT studies*. Journal of Physical Chemistry A, 2002. **106**(29): p. 6775-6782
13. Grabowski, S.J., *Ab initio calculations on conventional and unconventional hydrogen bonds - Study of the hydrogen bond strength*. Journal of Physical Chemistry A, 2001. **105**(47): p. 10739-10746.
14. Arunan, E., et al., *Definition of a Hydrogen Bond*. IUPAC task group, 2010.
15. Quinn, J.R., et al., *Does the A.T or G.C base-pair possess enhanced stability? Quantifying the effects of CH..... O interactions and secondary interactions on base-pair stability using a phenomenological analysis and ab initio calculations*. Journal of the American Chemical Society, 2007. **129**(4): p. 934-941.

5 Non-complementary Heteronaphthalene Associations

5.1 Introduction

In order to fully explore the proposed Het set of letters all possible associations need to be constructed and the TIE of each determined within the STRD molecular geometry constraints. In any viable alphabet all possible Watson-Crick type associations of letters other than those that are complementary must result in a repulsive interaction. In total 136 (Watson-Crick) possible associations of letters exist (see appendix A5). These associations can be put into mismatch categories based on the number of mismatches present (0-4). The number of mismatches that exist between any two molecules can be determined formally by calculating the weight of the XNOR product (the reverse of the XOR discussed in section 1.2) which gives the complementary Hamming distance ($\bar{\partial}$). A sample XNOR function calculation for each of the 5 complementary Hamming distance categories can be seen below (Fig. 5.1)

Mismatch Distances	$\bar{\partial}$	Mismatch Distances	$\bar{\partial}$
HetD = 0011		HetG = 0110	
<u>HetD*</u> = 1100		<u>HetH*</u> = 1000	
XNOR(Het[DD*]) = 0000		XNOR(Het[GH*]) = 0001	
$\bar{\partial}$ (Het[DD*])=0		$\bar{\partial}$ (Het[GH*]) = 1	
Mismatch Distances	$\bar{\partial}$	Mismatch Distances	$\bar{\partial}$
HetF = 0101		HetG* = 1001	
<u>HetG</u> = 0110		<u>HetH*</u> = 1000	
XNOR(Het[FG]) = 1100		XNOR(Het[G*H*]) = 1110	
$\bar{\partial}$ (Het[FG])=2		$\bar{\partial}$ (Het[G*H]) = 3	
Mismatch Distances	$\bar{\partial}$		
HetD = 0011			
<u>HetD</u> = 0011			
XNOR(Het[DD]) = 1111			
$\bar{\partial}$ (Het[DD])=4			

Figure 5.1 Sample XNOR calculation for each of the $\bar{\partial}$ categories

The number of pairs of each mismatch type can be calculated based on XNOR and $\bar{\delta}$ values and the different ways in which identical $\bar{\delta}$ values can arise from different XNOR values (see appendix A5).

In this chapter all 128 mismatching associations ($\bar{\delta} = 1, \bar{\delta} = 2, \bar{\delta} = 3, \bar{\delta} = 4$) will be explored. This in this first instance will be done using HF with a 6-31G* basis set (G03W software), STRD geometric constraints (Table 5.1).

Table 5.1 Standard (STRD) conditions

Position	Distance (Å)	Angle (Degrees)	Point Group
$\beta(2^{\text{nd}})$	3.14	180	C_s
$\gamma(3^{\text{rd}})$	3.14	180	

The model proposed by Mac Dónaill (see section 1.4) suggests that the patterns in the Het alphabet are inherently disadvantaged compared to alphabets such as the nucleotide. In the Het alphabet each pattern has a complementary pattern also contained within the alphabet. This complement requirement means that the maximum mutual minimum distance that can be achieved in a set of codewords with 4 D/A positions is $\delta = 2$. In the nucleotide alphabet, in which two distinct sizes are used a distance a larger relative distance can exist.

5.1.2 Imposing standard geometric constraints in mismatching positions

Two types of mismatch are possible, lone pair – lone pair (Lp-Lp) and hydrogen – hydrogen (H-H). STRD conditions will be applied to all associations even where mismatches occur in the middle two positions ($\beta(2^{\text{nd}})$ or $\gamma(3^{\text{rd}})$). In the case of a H-H mismatch (Fig. 5.2); the distance restriction (3.14Å) is still imposed from heavy atom to heavy atom (1-4 Fig. 5.2). In order to keep the pairs rigid at 180° angles it was decided to lock the angles from both (1-2-4 and 1-3-4 Fig. 5.2) sides to avoid introducing any bias and to keep associations directly comparable.

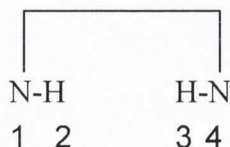


Figure 5.2 N-H... H-N mismatch. Distance constrained N-N. Angles constrained 1-2-4, 1-3-4.

If a Lp-Lp mismatch is present in one of the middle positions (Fig. 5.3) the distance is once again locked from heavy atom to heavy atom (2-3 Fig. 5.3). To enable the angle to be held from both sides carbon atoms from the ring structure must be used (1-2-4 and 1-3-4) Fig. 5.3).

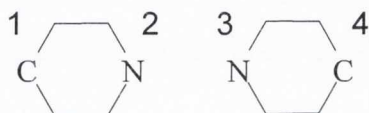


Figure 5.3 C...N...N...C mismatch. Distance constrained N-N. Angles constrained 1-2-4, 1-3-4.

All H-H or Lp-Lp mismatches occurring in a middle position will be constrained in the ways described above during the investigation of all non-complementary interactions.

5.2 Non-complementary Heteronaphthalene associations-Results

5.2.1 Mismatches in one position

32 associations that mismatch in one position exist (see appendix A5 for details of how this can be calculated). Each association has either a single Lp-Lp or H-H mismatch present (Fig. 5.4a, Fig. 5.4b).

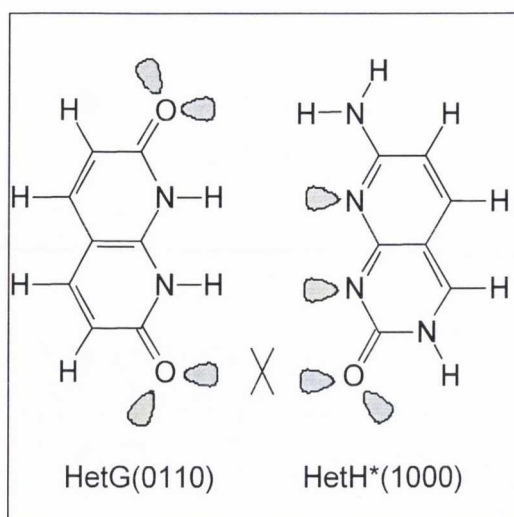


Figure 5.4a Het[GH*] Lp-Lp mismatch

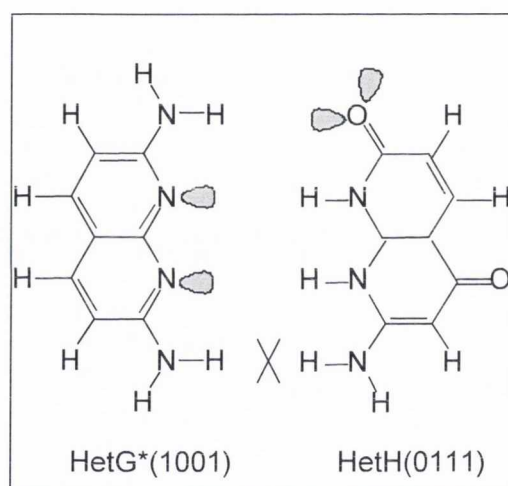


Figure 5.4b Het[G*H] H-H mismatch

In order to aid in the analysis of the 32 pairs, groupings can be made by XNOR function (which details the position in which a mismatch occurs), the type of mismatch present and by complementary letter. Grouping by complementary letter results in each pair being related to another, Het[GH*] 0110-1000 (as seen in Fig. 5.4a), for example, is related by complementary association to Het[G*H] 1001-0111 (Fig. 5.4b). In any viable set of letters each letter must have its complementary letter contained also within the set. This means that in order for Het[G*H] to be part of a viable set Het[G*H] must also meet all requirements and be part of the set. Any pair cannot be considered entirely independently from those it is related to by complementary association. Two pairs such as Het[GH*] and

Het[G*H] shall be referred to as a complementary couple. In each couple one pair will have a H-H mismatch and the other a Lp-Lp.

The TIE for each of the 32 associations was calculated using Eqn. 5.1.

$$E_{TIE}(AB) = E(AB) - (E(A) + E(B)) \quad \text{Equation 5.1}$$

The TIE of each mismatch in one position is shown in Table 5.2. In the data table (5.2) the results are arranged by XNOR function grouping and also by complementary couple (left to right).

Table 5.2 TIEs for all mismatches in one position broken into groups by the XNOR function and by complementary couple (from the left to the right side of the table)

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
Lp-Lp XNOR 1000		H-H XNOR 1000	
Het[BG] 0001-0110	-91.446	Het[B*G*] 1110-1001	-47.843
Het[AH] 0000-0111	-44.606	Het[A*H*] 1111-1000	-39.662
Het[DE] 0011-0100	-102.683	Het[D*E*] 1100-1011	1.296
Het[CF] 0010-0101	-36.194	Het[C*F*] 1101-1010	9.708
Lp-Lp XNOR 0100		H-H XNOR 0100	
Het[DH*] 0011-1000	-95.313	Het[D*H] 1100-0111	-25.842
Het[BF*] 0001-1010	-60.977	Het[B*F] 1110-0101	-3.394
Het[CG*] 0010-1001	-43.307	Het[C*G] 1101-0110	11.865
Het[AE*] 0000-1011	-29.706	Het[A*E] 1111-0100	23.596
Lp-Lp XNOR 0010		H-H XNOR 0010	
Het[BD*] 0001-1100	-97.910	Het[B*D] 1110-0011	-27.880
Het[FH*] 0101-1000	-62.149	Het[F*H] 1010-0111	-3.395
Het[EG*] 0100-1001	-43.054	Het[E*G] 1011-0110	11.875
Het[AC*] 0000-1101	-12.107	Het[A*C] 1111-0010	20.416
Lp-Lp XNOR 0001		H-H XNOR 0001	
Het[GH*] 0110-1000	-90.881	Het[G*H] 1001-0111	-47.844
Het[CD*] 0010-1100	-77.161	Het[C*D] 1101-0011	-20.217
Het[EF*] 0100-1010	-34.593	Het[E*F] 1011-0101	9.700
Het[AB*] 0000-1110	-32.001	Het[A*B] 1111-0001	-42.746

The results show that most of the associations are still binding (have a negative TIE). This indicates that a mismatch in just one out of a possible four positions is not always sufficient to prevent binding. A spread of energy values is evident (Het[DE] = -102.683 kJ/mol, Het[A*E] = 23.596 kJ/mol) suggesting that all mismatches are not of equal magnitude.

Examining the complementary couples, no example can be found where both pairs in the couple are repulsive. Het[A*C] for example has a TIE of 20.416 kJ/mol but the other association in its complementary couple Het[AC*] had an attractive energy of -12.107 kJ/mol.

The TIE results are also shown in the figure below (Fig. 5.5). In this plot it can be seen (by inspection) that in almost all cases a H-H mismatch gives a more repulsive interaction energy than a Lp-Lp.

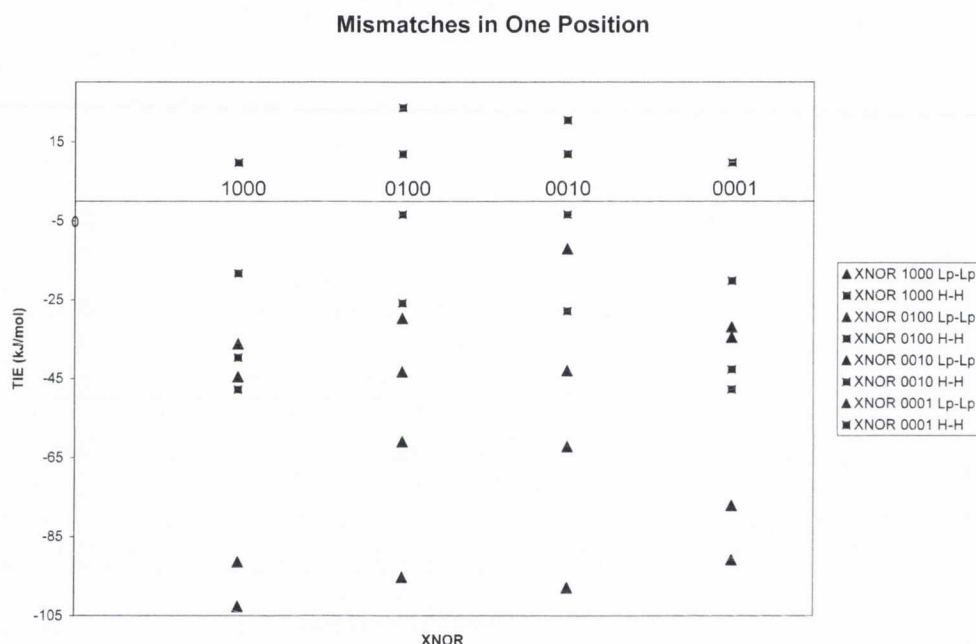


Figure 5.5 TIE for mismatches in one positions. Pairs grouped by type of mismatch and position.

The TIE values calculated for mismatches in the alpha position closely mirror the results of those in the delta position. The middle two positions β and γ also resemble each other in the pattern of interaction energies seen due to them being constrained in the same way and being equally close to a free terminal position. Essentially, a further grouping of mismatching association could be made by grouping α and δ together as terminal and β and γ as middle. This has not been done on the graph at this stage in order to facilitate clarity in the analysis of results.

Mismatches in the middle two positions, where STRD conditions are in place, show slightly higher repulsions compared to the terminal positions.

As noted in section 1.5 one of the conditions for a potentially viable alphabet is that each letter must have a complementary letter which is contained within the alphabet; Het[AC*] and Het[A*C] cannot be considered completely independently of each other. All of the Lp-Lp mismatches are binding and as each of these (Lp-Lp) association is related to a H-H they are termed the limiting associations. Each Lp-Lp mismatching association prevents a related H-H association (no matter how repulsive it may be) from remaining in a viable alphabet. In the example given above, both Het[GH*] and Het[G*H] must be removed from the alphabet even though Het[A*C] is repulsive as it is limited by Het[AC*] which is binding. Applying this logic to all associations that mismatch in one position, a Het alphabet that contains pairs that mismatch in only one position cannot exist.

A single mismatch is not enough to ensure non-binding. This finding agrees with that published by Mac Dónaill and Brocklebank [1] for the nucleotide alphabet.

5.2.2 Mismatches in two positions

48 associations can be formed that mismatch in two out of four positions (see appendix A5 for calculation details). These associations can contain two Lp-Lp mismatches, two H-H or one mismatch of each type (Fig. 5.6).

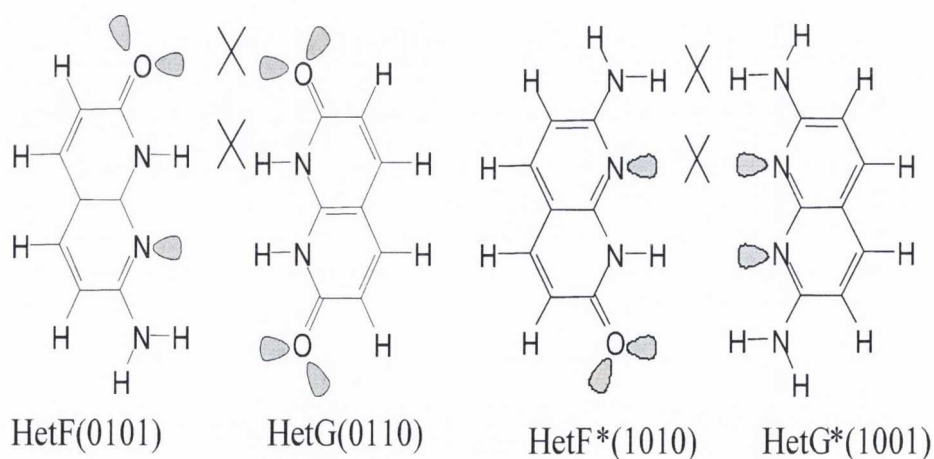


Figure 5.6 Examples of pairs with two mismatches.

This group of mismatches can be divided into sets based on XNOR function, the type of mismatch present (H-H or Lp-Lp) and complementary couple and parity (section 1.4). Letters that mismatch in two positions only occur where both letters have the same parity.

It is therefore convenient to consider interactions where both letters have odd or even parity. {A,A*,D,D*,F,F*,G,G*} are even (contain an even number of 1s in each codeword) and {B,B*,C,C*,E,E*,H,H*} are odd.

The model proposed by Mac Dónaill (section 1.4) sets out that for a set of 4-bit D/A patterns the maximum mutual distance (δ) that can be achieved is 2. In an alphabet (such as the Het) where all associations are considered in one group which requires complements, the $\delta=2$ set of associations should have the greatest chance of containing a potentially viable subset.

On inspection of the interaction energy results (Table 5.3a, 5.3b), many of the associations are repulsive but a few associations (of each parity) are binding. In order to form a viable subset of the Het potential alphabet letter set all attractive associations (and any related to them by complement) need to be ruled out from remaining in a potential subset. Five sets of associations each containing two complementary couples exist in which all associations are repulsive (the five sets are highlighted below in black).

Table 5.3a TIEs for all even parity pairs with two mismatches

Pair Even	TIE(kJ/mol)	Pair Even	TIE(kJ/mol)
XNOR 1100		XNOR 0101	
Het[AD] 0000-0011	-22.730	Het[DG*] 0011-1001	38.960
Het[A*D*] 1111-1100	120.041	Het[D*G] 1100-0110	41.572
XNOR 0011		XNOR 1010	
Het[AD*] 0000-1100	4.550	Het[D*G*] 1100-1001	37.624
Het[A*D] 1111-0011	116.310	Het[DG] 0011-0110	40.431
XNOR 0011		XNOR 0110	
Het[FG*] 0101-1001	29.528	Het[AG*] 0000-1001	40.894
Het[F*G] 1010-0110	52.188	Het[A*G] 1111-0110	144.832
XNOR 1100		XNOR 1001	
Het[F*G*] 1010-1001	29.528	Het[AG] 0000-0110	-4.379
Het[FG] 0101-0110	52.187	Het[A*G*] 1111-1001	77.844
XNOR 0101		XNOR 0110	
Het[AF*] 0000-1010	7.583	Het[D*F] 1100-0101	34.334
Het[A*F] 1111-0101	107.051	Het[DF*] 0011-1010	34.741
XNOR 1010		XNOR 1001	
Het[AF] 0000-0101	12.099	Het[DF] 0011-0101	45.968
Het[A*F*] 1111-1010	107.052	Het[D*F*] 1100-1010	46.770

Table 5.3b TIEs for all odd parity pairs with two mismatches

Pair Odd	TIE(kJ/mol)	Pair Odd	TIE(kJ/mol)
XNOR 1100		XNOR 0101	
Het[BC] 0001-0010	-6.665	Het[CH*] 0010-1000	-23.231
Het[B*C*] 1110-1101	117.027	Het[C*H] 1101-0111	86.287
XNOR 0011		XNOR 1010	
Het[BC*] 0001-1101	7.644	Het[C*H*] 1101-1000	56.763
Het[B*C] 1110-0010	35.689	Het[CH] 0010-0111	60.133
XNOR 0011		XNOR 0110	
Het[EH*] 0100-1000	-35.619	Het[BH*] 0001-1000	12.739
Het[E*H] 1011-0111	117.075	Het[B*H] 1110-0111	110.285
XNOR 1100		XNOR 1001	
Het[E*H*] 1011-1000	8.771	Het[BH] 0001-0111	4.094
Het[EH] 0100-0111	11.194	Het[B*H*] 1110-1000	7.675
XNOR 0101		XNOR 0110	
Het[B*E] 1110-0100	36.108	Het[C*E] 1101-0100	39.481
Het[BE*] 0001-1011	62.871	Het[CE*] 0010-1011	65.087
XNOR 1010		XNOR 1001	
Het[BE] 0001-0100	-51.890	Het[CE] 0010-0100	-43.142
Het[B*E*] 1110-1011	86.288	Het[C*E*] 1101-1011	89.125

Out of the five subsets in which all associations (both complementary couples) are repulsive, two of the sets contain at least one association which is only very weakly repulsive. The set of pairs for letters HetA and HetF contains the association Het[AF] which only has a repulsive energy of 7.583 kJ/mol. The STRD conditions used in this exploration are likely to overestimate the repulsions between letters. If the pairs were given greater freedom of movement they would most likely be able to use this freedom to somewhat relieve repulsions. As this is the case a repulsion of 7.583 kJ/mol is unlikely to be sufficient to prevent binding under different conditions and would need to be removed from the alphabet. Due to its weak repulsion Het[AF] can be described as the limiting pair for the group of possible association between HetA and HetF. The low repulsion of Het[AF] leads to the removal of the entire HetA and HetF set from any potential alphabet as even though three out of the four non-complementary associations are in fact repulsive they are limited by the weakest link Het[AF].

The only potentially viable set of complementary couples in the odd parity letter group is that formed between HetB and HetH. This set is limited by its weakest interaction Het[AH] which has a TIE of only 4.084 kJ/mol. This as was the case for Het[AH] is unlikely to be repulsive enough to prevent binding, particularly under constraints less rigid than the STRD.

In the three remaining sets of complementary couples all interactions are strongly repulsive. These three set contain six letters HetD, HetD, HetF, HetF*, HetG and HetG*. These three remaining sets are all independently potentially viable but as all possible associations between these six letters (other then those that are complementary or self mismatching) are contain two mismatches they can be considered as a larger potentially viable letter set {Het[DD*], Het[FF*], Het[GG*]}.

The results can be visualised on the plot shown below (Fig. 5.7). Similar to mismatches in one position, associations with two mismatches are grouped by XNOR value and the type of mismatch present.

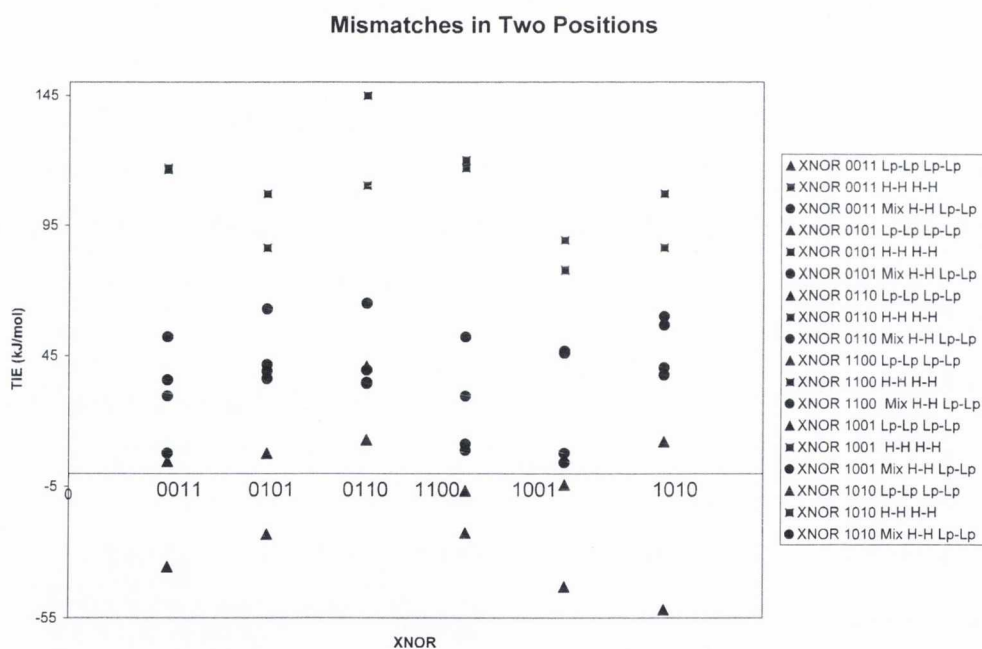


Figure 5.7 TIE for mismatches in two positions. Pairs grouped by XNOR and mismatch type.

The graph highlights that it is the pairs with two H-H mismatches which have the most repulsive energy and those with two Lp-Lp the least. A mismatch in one of the constrained middle two positions causes the pairs to be more repulsive particularly in the case of H-H repulsions. The limiting pairs (those that are still binding) all contain two Lp-Lp mismatches.

The entire set of associations which mismatches in two out of a possible four positions cannot exist together. Even at the maximum mutual distance theoretically possible for 4-bit patterns non-appropriate matches still occur. A subset of letters can be seen however in which all interactions other than those that are complementary are repulsive enough to prevent binding. This overall result differs from that seen for nucleotides in which all eight even parity pairs survived based on interaction energy alone. Six out of the eight potentially viable pairs were knocked out due to chemical limitations [1].

5.2.3 Mismatches in three positions

The 32 pairs that mismatch in three positions (Fig. 5.8) can contain any one of four mismatch type combinations;

1. Lp-Lp Lp-Lp Lp-Lp
2. H-H H-H H-H
3. Lp-Lp Lp-Lp H-H
4. H-H H-H Lp-Lp

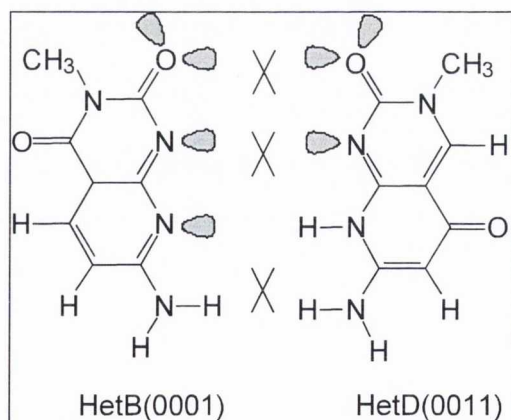


Figure 5.8 Pair with three mismatches Het[BD]

By inspection of the results it can be seen that all pairs with three mismatches are repulsive (Table 5.4).

Table 5.4 TIE for mismatches in three positions

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
XNOR 0111		XNOR 1101	
Het[AH*] 0000-1000	62.675	Het[AC] 0000-0010	22.448
Het[A*H] 1111-0111	243.242	Het[A*C*] 1111-1101	224.749
Het[DE*] 0011-1011	177.516	Het[EG] 0100-0110	65.727
Het[D*E] 1100-0100	70.235	Het[E*G*] 1011-1001	129.925
Het[CF*] 0010-1010	93.246	Het[BD] 0001-0011	91.439
Het[C*F] 1101-0101	131.486	Het[B*D*] 1110-1100	169.775
Het[BG*] 0001-1001	109.07	Het[FH] 0101-0111	142.992
Het[B*G] 1110-0110	172.189	Het[F*H*] 1010-1000	72.632
XNOR 1011		XNOR 1110	
Het[AE] 0000-0100	10.491	Het[AB] 0000-0001	54.689
Het[A*E*] 1111-1011	224.745	Het[A*B*] 1111-1110	243.238
Het[BF] 0001-0101	71.877	Het[EF] 0100-0101	66.464
Het[B*F*] 1110-1010	142.684	Het[E*F*] 1011-1010	131.477
Het[DH] 0011-0111	166.602	Het[CD] 0010-0011	97.03
Het[D*H*] 1100-1000	90.199	Het[C*D*] 1101-1100	178.201
Het[CG] 0010-0110	91.013	Het[GH] 0110-0111	172.189
Het[C*G*] 1101-1001	129.919	Het[G*H*] 1001-1000	108.434

Studying the results graphically (Fig 5.9) reveals that the greater the number of H-H mismatches present the higher the TIE repulsion. Although the remaining mismatches are all repulsive they could not be used to form a viable subset of the alphabet as each pair is related to a mismatch in one position through complementary coupling. If any one of the letters in a given pair that mismatches in three positions is changed to its complement (Eg, Het[BD]-Het[BD*]), a mismatch in one position will be created. Due to this relationship existing between three and one mismatches, associations that mismatch in three positions must also be ruled out from forming part of a viable alphabet as mismatches in one position limit their potential viability.

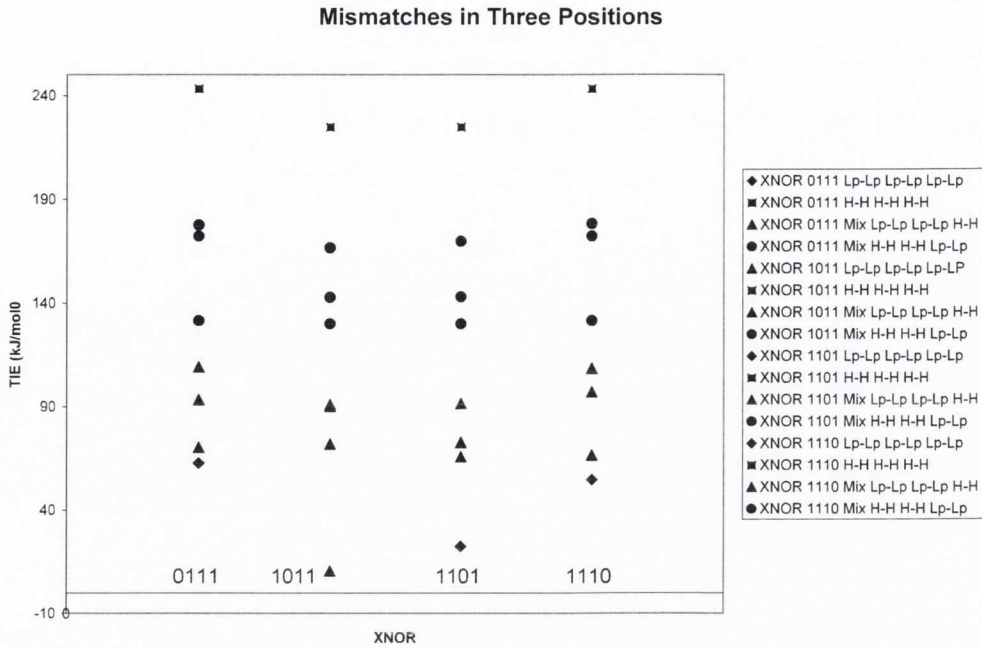


Figure 5.9 TIE for mismatches in three positions. Pairs grouped by XNOR value and type of mismatch.

The relationship that exists between one and three mismatches is the primary inherent disadvantage of a 4-bit alphabet which must contain complements and consider all letters in one group. As discussed in section 1.4 for a set of 4-bit patterns the maximum mutual separation that can be achieved is $\delta = 2$. In an alphabet like the Het creating a set (that contains complements) of associations that has a separation greater than 2 implies that the set must also contain associations with less than two mismatches (Eqn. 5.2). As the number of mismatches increases between two letters ab , it decreases between ab^* . The total number of mismatches in a complementary couple must always be equal to the number of bits (n). If each complementary association in the Het alphabet was partitioned in some way that prevented all associations from mixing (for example by size), the relationship between one and three mismatches could be overcome and a potentially viable set of letters each containing three mismatches could be formed.

$$\partial(a,b) + \partial(a,b^*) = n \qquad \text{Equation 5.2}$$

5.2.4 Mismatches in four positions

In order to complete the exploration of the Het set of molecules mismatches in four positions must also be considered (Fig. 5.10). Mismatches of this type are formed by pairing each of the 16 letters with itself.

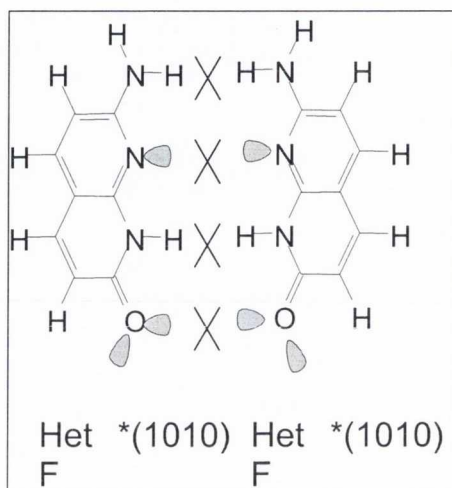


Figure 5.10 Het[F*F*] with four mismatches.

All of these pairings give a repulsive TIE (Table 5.5). In the analysis of this set of associations pairs are not grouped by XNOR value (as in all cases it is 1111) but by the number of H-H mismatches present (Fig. 5.11). A strong correlation between the number of H-H mismatches and the TIE is present. A unique range of values is seen for each set of pairs with a given number of H-H interactions present.

Table 5.5 TIEs for mismatches in four positions

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
Het[AA] 0000-0000	86.756	Het[A*A*] 1111-1111	385.747
Het[BB] 0001-0001	151.571	Het[B*B*] 1110-1110	285.200
Het[CC] 0010-0010	134.624	Het[C*C*] 1101-1101	258.720
Het[DD] 0011-0011	223.871	Het[D*D*] 1100-1100	224.297
Het[EE] 0100-0100	140.055	Het[E*E*] 1011-1011	258.720
Het[FF] 0101-0101	182.290	Het[F*F*] 1010-1010	182.290
Het[GG] 0110-0110	217.341	Het[G*G*] 1001-1001	190.479
Het[HH] 0111-0111	285.203	Het[H*H*] 1000-1000	149.867

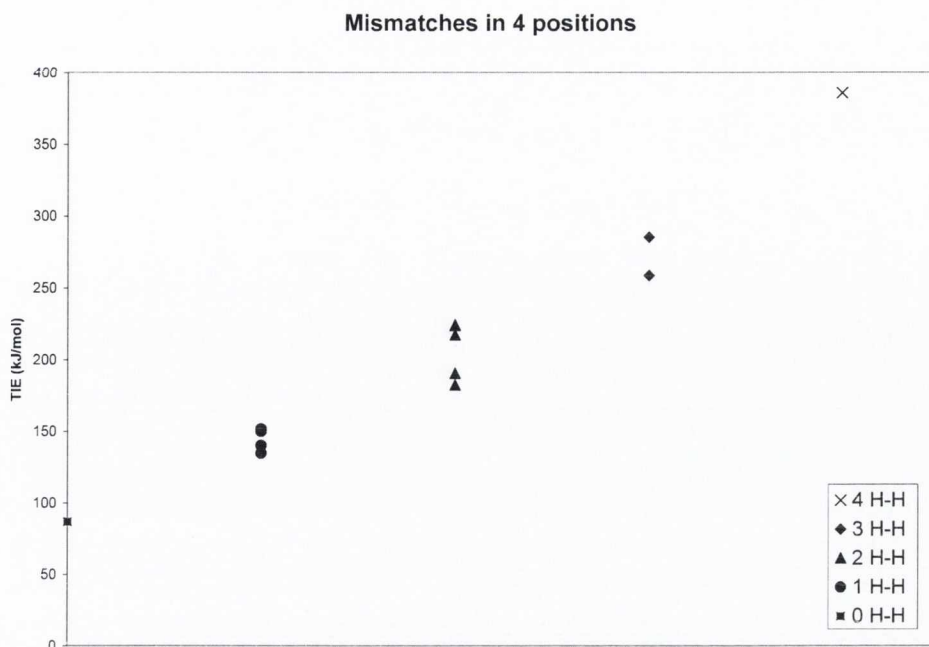


Figure 5.11 TIE for mismatches in four positions. Pairs grouped by number of H-H mismatches

5.3 Heteronaphthalene HF results: Discussion and Conclusions

Having carried out an exploration of all possible associations for the Het alphabet the results reveal that it is not possible for all 16 molecules to coexist. If they were to coexist binding pairs other than those that are complementary could be formed.

In order for a potential alphabet to meet the interaction energy criteria for viability each molecule needs to bind strongly to its complementary letter and to repel all other letters. In the Het alphabet many non-complementary associations remain binding (indicated by a negative TIE). Thus, in order to ensure binding fidelity the possibility of forming any of these binding non-complementary associations must be removed from the potential alphabet. This is done by removing letters from the potential set thus reducing the size of the alphabet. Every mismatching association is related (by complementary letters) to three others. An example of a complete set of four mismatching associations for Het[B], Het[B*], Het[D], Het[D*] can be seen below (Table 5.6). In order for Het[BD] to be part of a viable alphabet (or subset of an alphabet) it must be repulsive but so must the other three associations related to it.

Table 5.6 Pair set for Het[B], Het[B*], Het[D], Het[D*]

Het[BD] 0001-0011	Het[B*D*] 1110-1100
XNOR 1101	XNOR 1101
Het[BD*] 0001-1100	Het[B*D] 1110-0011
XNOR 0010	XNOR 0010

Mismatches in one position are related to those in three positions as described by (Eqn 1.4). In the pair set shown above (Table 5.6) two of the pairs, Het[BD*] and Het[B*D], mismatch in one position whilst the other two pairs in the set, Het[BD] and Het[B*D*], mismatch in three positions. All mismatches in one and three positions are related in this way and thus cannot be considered independently. No subset of pairs that mismatch in one position can be found that meets the energy criteria for viability. All mismatches in one position must be ruled out and because of complementary association so must all mismatches in three positions. The fact that associations which mismatch in one and in three positions are related in this way limits the potential for viability of the Het alphabet.

As proposed by Mac Dónaill, the maximum mutual distance that can be achieved for a 4-bit alphabet is $\delta=2$. If the complementary associations in the Het alphabet were to be split into groups, by size for example, a significant advantage would be gained as mismatches in one and three positions could be separated. Mismatches in two positions on the other hand are related only to other mismatches in two positions by complementary association (Fig. 5.7). All pairings in a given set are equidistant from one another ($\hat{\sigma}=2$).

Table 5.7 Pair set for Het[D], Het[D*], Het[G], Het[G*]

Het[DG] 0011-0110	Het[D*G*] 1100-1001
XNOR 1010	XNOR 1010
Het[DG*] 0011-1001	Het[D*G] 1100-0110
XNOR 0101	XNOR 0101

In the case of mismatches in two positions a subset of 6 letters can be found in which every possible combination (other than those that are complementary) results in a repulsive interaction energy (Table 5.8).

Table 5.8 TIE in kJ/mol Het[DD*], Het[FF*] Het[GG*], surviving pairs

Pair	TIE (kJ/mol)
XNOR 0011	
Het[FG*] 0101-1001	29.528
Het[F*G] 1010-0110	52.188
XNOR 1100	
Het[F*G*] 1010-1001	29.528
Het[FG] 0101-0110	52.187
XNOR 0101	
Het[DG*] 0011-1001	38.960
Het[D*G] 1100-0110	41.572
XNOR 1010	
Het[D*G*] 1100-1001	37.624
Het[DG] 0011-0110	40.431
XNOR 0110	
Het[D*F] 1100-0101	34.334
Het[DF*] 0011-1010	34.741
XNOR 1001	
Het[DF] 0011-0101	45.968
Het[D*F*] 1100-1010	46.770

Although these pairs do meet the necessary energy requirements for viability they must also meet all of the other conditions including chemical before being classed as a true surviving subgroup.

Considering the secondary interactions present in mismatching associations does not aid in describing (or predicting) the order of TIEs. Many of pairs have an equal number of net secondary interactions (SI), within these groups (of equal net secondary interactions) a wide variation is still evident (Table 5.9). This variation cannot be addressed in a similar way to that of complementary associations in which a weighted fitting equation was designed. The equation designed in section 4.2.1 was designed based not purely on the number of SI present but also on the positions in which the interaction occurred. For mismatching associations, as in the sample set shown below, many of the pairs are not only equal in net SI but also in middle block (Fig. 5.12) interaction terms ruling out the possibility of a weighted approach.

Table 5.9 Pairs with equal numbers of mismatches and net secondary interactions

	TIE (kJ/mol)	Net. SI
XNOR 0101 Lp-Lp Lp-Lp		
Het[AF*] 0000-1010	7.583	0
Het[CH*] 0010-1000	-23.231	0
XNOR 0101 H-H H-H		
Het[A*F] 1111-0101	107.051	0
Het[C*H] 1101-0111	86.287	0
XNOR 0101 Mix Lp-Lp H-H		
Het[B*E] 1110-0100	36.108	0
Het[BE*] 0001-1011	62.871	0
Het[DG*] 0011-1001	38.960	0
Het[D*G] 1100-0110	41.572	0

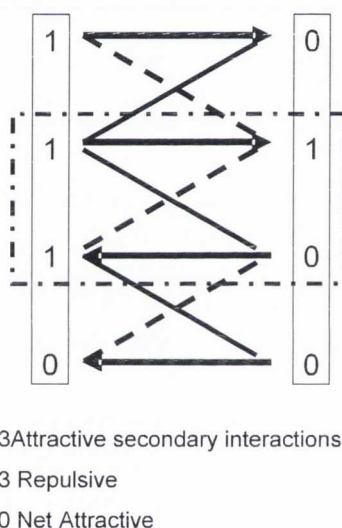


Figure 5.12 Example Het[B*E] with 0 middle secondary interactions

To see the overall trend in results a plot can be made of the total interaction energy for each pair versus the number of mismatches (Fig. 5.13). As the number of mismatches increases, the total interaction energy becomes more repulsive with fewer remaining negative points.

A large spread of energies can be seen at each point on the x-axis. The spread of total interaction energy values for a given number of mismatches is not unique to that number of mismatches. Het[AE] (TIE=10.491 kJ/mol), for example, which mismatches in two positions has a less repulsive interaction energy than Het[A*C] (TIE= 20.416kJ/mol) which mismatches in only one position.

The spread in interaction energies may be slightly overestimated as the total interaction energies for each distinct mismatch set can be broken into groups of pairs related by complement. In doing this each mismatch in one position is related to one other mismatch in one position. Het[AH], for example, is related to Het[A*H*] and these cannot occur independently of each other. The same grouping procedure applies to mismatches in three positions. Mismatches in two positions can be grouped into sets of four pairs, as all possible groupings by complement in this case will mismatch in two positions. The lowest, most attractive total interaction energy for each pair in a given letter group/set can be described as the limiting pair. By considering only the limiting pairs from each group the data set can be reduced (Fig. 5.14).

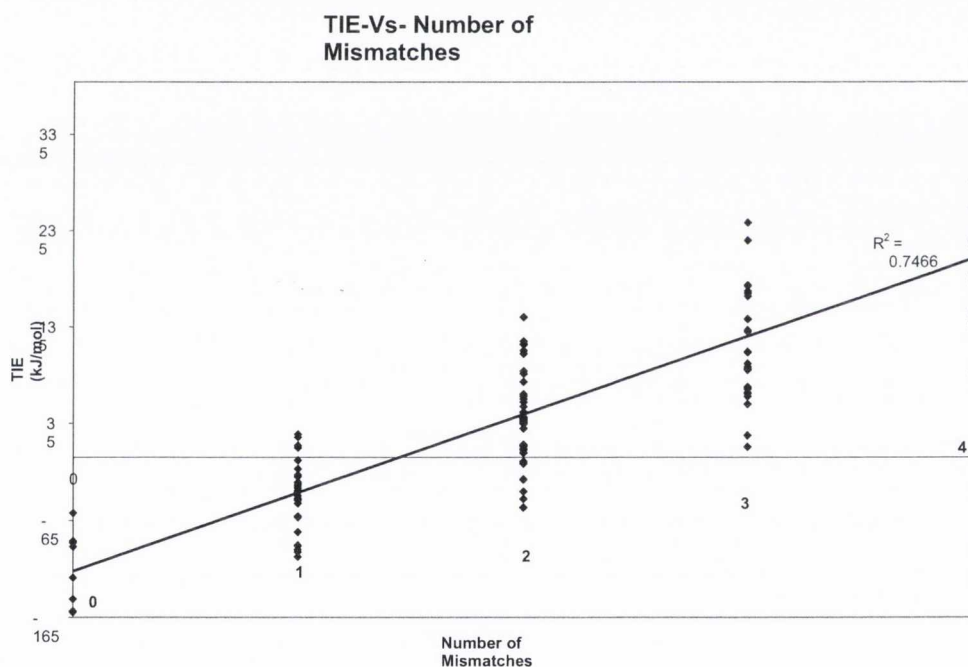


Figure 5.13 TIE versus Number of mismatches for all Het associations

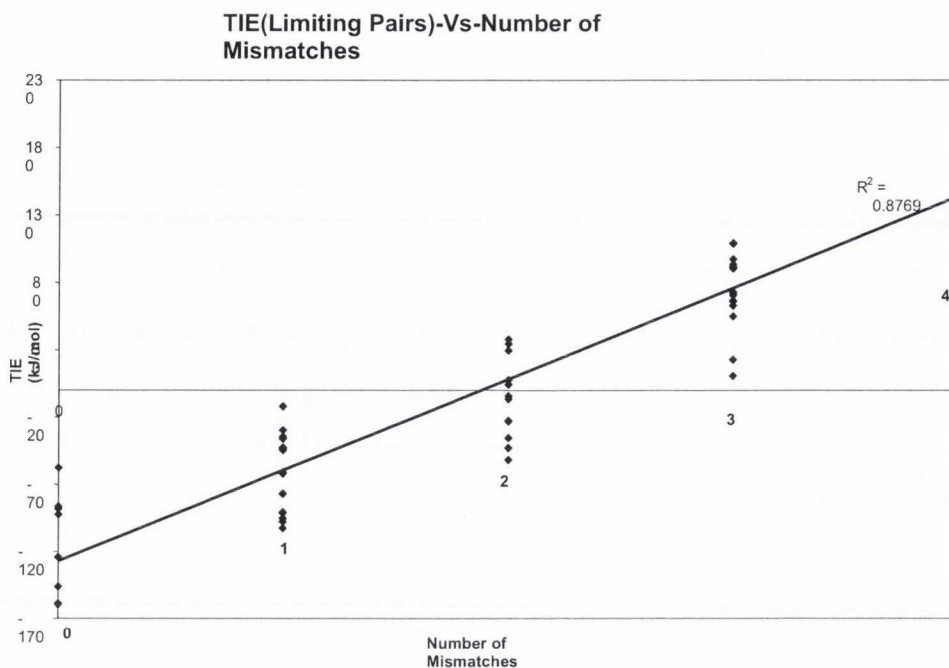


Figure 5.14 TIE versus number of mismatches for limiting pairs only.

Here again, as was the case for the complete set of pairs, it can be seen that the spread of total interaction energies for a given number of mismatches is not unique to that number of mismatches. These results differs to those seen for the nucleotide alphabet in which no overlap between the number of mismatches remains when only limiting associations are considered [1] (Fig. 5.15).

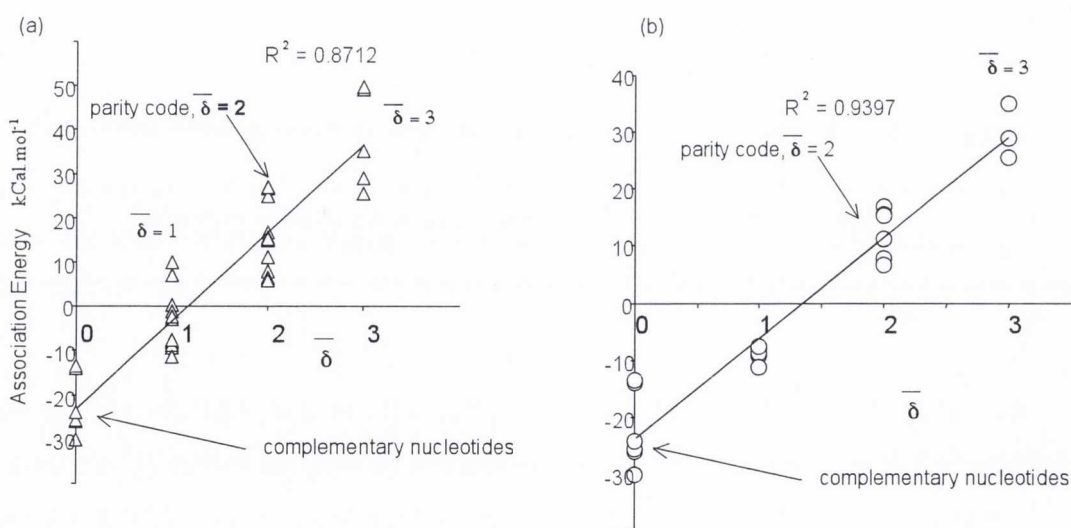


Figure 5.15 (a) all data for nucleotide associations (b) limiting data nucleotide associations. Diagram taken directly from [1]

Although the Het alphabet is inherently disadvantaged as its structure leads to the unavoidable relationship between one and three mismatches, the possibility of a viable subset of letters at the maximum mutual distance of $\delta=2$ cannot be ruled out.

In the next chapter we address the possibility of artefacts arising from the computational methods employed.

1. MacDonaill, D.A. and D. Brocklebank, *An ab initio quantum chemical investigation of the error-coding model of nucleotide alphabet composition*. Molecular Physics, 2003. **101**(17): p. 2755-2762.

6 BSSE and Different Computational Methods

6.1 Introduction

The results in the previous chapter are of course subject to the suitability of HF calculations and ignore well known deficiencies such as the Basis Set Superposition Error (BSSE) and the absence of correlation. Accordingly in this chapter we repeat calculations using a variety of methodologies with a view to confirming that the conclusions from the behaviour of the Het alphabet are not artefacts. A variety of computational approximations were selected, ranging from the semi-empirical (AM1, PM3) to *ab initio* Hartree-Fock with MP2, so as to minimise the possibility of artefacts arising from the characteristics of any one method. Alphabet properties which are essentially common across the various computational approximations are considered likely to be reliable.

6.2 Basis Set Superposition Error

Basis set superposition error (BSSE) which arises due to the inconsistency in basis set size between dimer and monomers (section 2.4) is present in each of the TIEs calculated in the previous chapter. Without the removal of BSSE each interaction energy will be artificially more binding. In order to assess the overall effect of BSSE on the Het potential alphabet it will be calculated and removed. In this section the influence of BSSE will first be explored for complementary and mismatching associations. The BSSE removal calculations will be carried out at the same computational level (HF) and with the same basis set (6-31G*) as before. BSSE will be eliminated from each dimer through a single point (SP) energy calculation that includes a counterpoise (CP) correction (section 2.4).

6.3 BSSE calculation results

6.3.1 BSSE complementary associations

SP calculations including CP correction were carried out on all 8 previous optimized complementary associations with STRD geometry restrictions in place. A plot can be made summarizing the results of the contribution of BSSE to each TIE (Table 6.1) (Fig. 6.1).

Table 6.1 TIEs for complementary associations with and without BSSE correction

Pair	TIE (kJ/mol)	CP Corrected TIE (kJ/mol)	BSSE (kJ/mol)
Het[AA*] 0000-1111	-57.718	-46.739	-10.979
Het[BB*] 0001-1110	-159.61	-144.716	-14.894
Het[CC*] 0010-0010	-88.338	-73.778	-14.560
Het[DD*] 0011-1100	-146.232	-132.318	-13.914
Het[EE*] 0100-1011	-92.378	-77.632	-14.746
Het[FF*] 0101-1010	-86.551	-69.749	-16.803
Het[GG*] 0110-1001	-124.236	-107.674	-16.562
Het[HH*] 0111-1000	-158.645	-143.693	-14.952

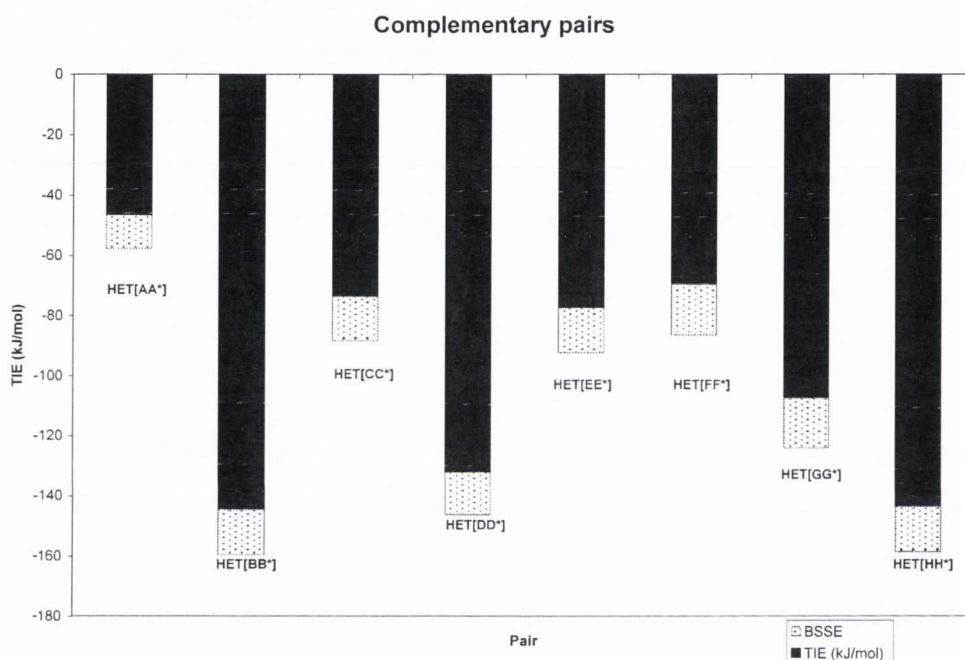


Figure 6.1 TIE for STRD complementary pairs showing the contribution of BSSE.

As expected on removal of BSSE each association becomes less binding. The BSSE contribution varies between -10.979 and -16.803 kJ/mol across the matching pairs. It is important to note that although the exclusion of BSSE does change the absolute value of each TIEs it does not change the relative ordering of energy values.

6.3.2 BSSE mismatches in one and three positions

As mismatches in one and three positions cannot be viewed entirely separately from each other (discussed in section 5.2.3), the effect of BSSE on these mismatch categories will be

considered together. The results for the removal of BSSE from pairs that mismatch in one position can be seen in (Table 6.2) (Fig. 6.2). In the plot below pairs have been grouped into sets based on complementary letter, for example, Het[AH] is grouped with Het[A*H*]. The 32 mismatching associations can be broken into 16 complement pair sets, one pair in each set will have a Lp-Lp mismatch and the other pair a H-H.

Table 6.2 TIEs for pairs with one mismatch, with and without BSSE

	Pair	TIE (kJ/mol)	CP Corrected TIE (kJ/mol)	BSSE (kJ/mol)
XNOR 1000 Lp-Lp	Het[AH] 0000-0111	-44.606	-32.696	-11.911
	Het[BG] 0001-0110	-91.446	-76.915	-14.530
	Het[DE] 0011-0100	-102.683	-89.558	-13.125
	Het[CF] 0010-0101	-36.194	-21.777	-14.416
XNOR 0100 Lp-Lp	Het[AE*] 0000-1011	-29.706	-18.269	-11.437
	Het[BF*] 0001-1010	-60.977	-45.440	-15.537
	Het[CG*] 0010-1001	-43.307	-28.208	-15.099
	Het[DH*] 0011-1000	-95.313	-81.092	-14.222
XNOR 0010 Lp-Lp	Het[AC*] 0000-1101	-12.107	-0.260	-11.846
	Het[BD*] 0001-1100	-97.910	-83.653	-14.257
	Het[EG*] 0100-1001	-43.054	-27.750	-15.304
	Het[FH*] 0101-1000	-62.149	-46.465	-15.684
XNOR 0001 Lp-Lp	Het[AB*] 0000-1110	-32.001	-20.693	-11.308
	Het[CD*] 0010-1100	-77.161	-64.201	-12.960
	Het[EF*] 0100-1010	-34.593	-20.288	-14.306
	Het[GH*] 0110-1000	-90.881	-76.515	-14.366
XNOR 1000 H-H	Het[H*A*] 1000-1111	-39.662	-27.133	-12.530
	Het[B*G*] 1110-1001	-47.843	-33.251	-14.593
	Het[D*E*] 11-1011	-18.296	-5.424	-12.871
	Het[C*F*] 1101-1010	9.708	24.065	-14.357
XNOR 0100 H-H	Het[A*E] 1111-0100	23.596	37.049	-13.453
	Het[B*F] 1110-0101	-3.394	12.344	-15.738
	Het[C*G] 1101-0110	11.865	27.330	-15.464
	Het[D*H] 1100-0111	-25.842	-11.642	-14.200
XNOR 0010 H-H	Het[A*C] 1111-0010	20.416	33.640	-13.224
	Het[B*D] 1110-0011	-27.880	-13.615	-14.265
	Het[E*G] 1011-0110	11.875	27.346	-15.471
	Het[F*H] 1010-0111	-3.395	12.337	-15.731
XNOR 0001 H-H	Het[A*B] 1111-0001	-42.746	-30.177	-12.569
	Het[C*D] 1101-0011	-20.217	-7.206	-13.011
	Het[E*F] 1011-0101	9.700	24.047	-14.347
	Het[G*H] 1001-0111	-47.844	-33.251	-14.593

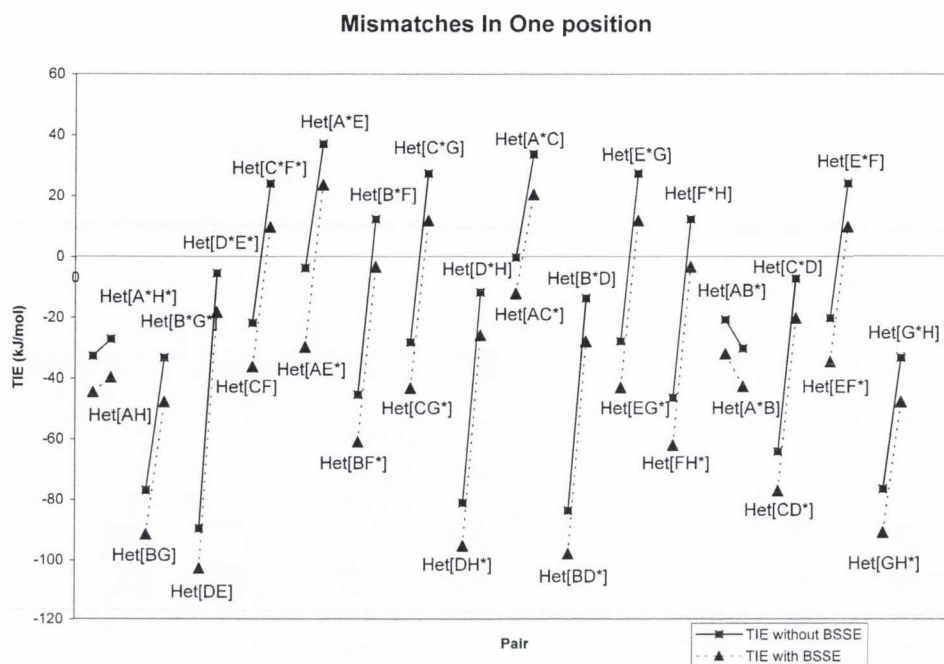


Figure 6.2 TIE for pairs that mismatch in one position with and without BSSE.

The results show that on removal of BSSE all of the associations become less binding (thus less negative) on average by -13.960 kJ/mol. On examination of the result plot (Fig. 6.2) it is apparent that although each energy has become less binding an overall shift in energies has taken place rather than a reordering of relative energies.

The results for mismatches in three positions present a similar picture to that seen for mismatches in one position (Table 6.3)(Fig. 6.3).

Table 6.3 TIEs for pairs with three mismatches, with and without BSSE

	Pair	TIE(kJ/mol)	CP Corrected TIE (kJ/mol)	BSSE (kJ/mol)
XNOR 1110	Het[AB] 0000-0001	54.689	65.406	-10.717
XNOR 1110	Het[A*B*] 1111-1110	243.238	255.331	-12.093
XNOR 1110	Het[CD] 0010-0011	97.030	109.516	-12.487
XNOR 1110	Het[C*D*] 1101-1100	178.201	190.759	-12.558
XNOR 1110	Het[EF] 0100-0101	66.464	80.828	-14.364
XNOR 1110	Het[E*F*] 1011-1010	131.477	145.628	-14.151
XNOR 1110	Het[GH] 0110-0111	172.189	186.397	-14.208
XNOR 1110	Het[G*H*]1001-1000	108.434	121.966	-13.531
XNOR 1101	Het[AC] 0000-0010	22.448	33.112	-10.664
XNOR 1101	Het[A*C*] 1111-1101	224.749	235.871	-11.121
XNOR 1101	Het[BD] 0001-0011	91.439	102.515	-11.076
XNOR 1101	Het[B*D*]1110-1100	169.775	181.892	-12.117
XNOR 1101	Het[EG] 0100-0110	65.727	79.062	-13.335
XNOR 1101	Het[E*G*] 1011-1001	129.925	142.570	-12.644
XNOR 1101	Het[FH] 0101-0111	142.992	156.400	-13.408
XNOR 1101	Het[F*H*]1010-1000	72.632	85.956	-13.324
XNOR 1011	Het[AE] 0000-0100	10.491	21.069	-10.577
XNOR 1011	Het[A*E*] 1111-1011	224.745	235.867	-11.123
XNOR 1011	Het[DH] 0011-0111	166.602	178.853	-12.251
XNOR 1011	Het[D*H*] 1100-1000	90.199	101.099	-10.900
XNOR 1011	Het[FB] 0101-0001	71.877	85.320	-13.443
XNOR 1011	Het[F*B*] 1010-1110	-3.390	12.344	-15.735
XNOR 1011	Het[CG] 0010-0110	91.013	104.444	-13.430
XNOR 1011	Het[C*G*] 1101-1001	129.919	142.560	-12.641
XNOR 0111	Het[AH*] 0000-1000	62.675	72.706	-10.031
XNOR 0111	Het[A*H] 1111-0111	243.242	255.340	-12.099
XNOR 0111	Het[BG*] 0001-1001	109.070	122.554	-13.484
XNOR 0111	Het[B*G] 1110-0110	172.189	186.398	-14.208
XNOR 0111	Het[CF*] 0010-1010	93.246	107.586	-14.340
XNOR 0111	Het[C*F] 1101-0101	131.486	145.645	-14.159
XNOR 0111	Het[DE*] 0011-1011	177.516	190.125	-12.609
XNOR 0111	Het[D*E] 1100-0100	70.235	82.480	-12.246

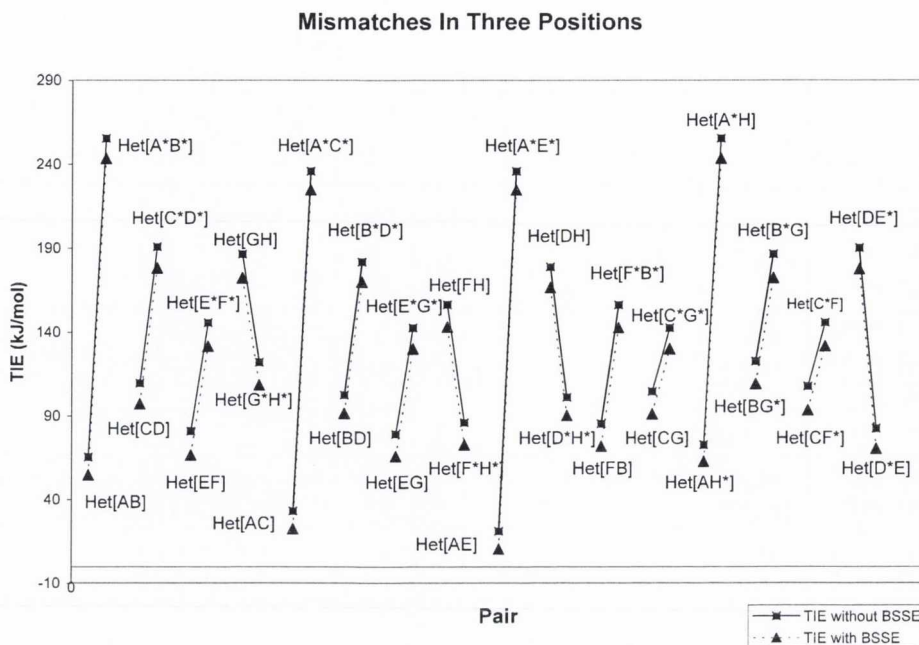


Figure 6.3 TIEs for pairs with three mismatches, with and without BSSE.

BSSE accounts for -12.659 kJ/mol on average of each TIE. The plot again reveals that although a shift in energy values is evident on elimination of BSSE the overall pattern of interaction energies remains unchanged.

Taking the effects of BSSE into account and removing it does not change the overall outcome of the study of pairs containing one or three mismatches. The absolute interaction energy values do change but the overall pattern of results remains unchanged.

6.3.3 BSSE mismatch in two positions

BSSE was removed from mismatches in two positions in the same manner as described above. The results are presented below in (Table 6.4) (Fig. 6.4). For convenience the results have been broken by complementary association into 8 subgroups each containing 4 pairs.

Table 6.4 TIEs for pairs with two mismatches, with and without BSSE

No. Key	Pair	TIE(kJ/mol)	CP Corrected TIE(kJ/mol)	BSSE (kJ/mol)
1	Het[AD] 0000-0011	-22.730	-11.796	-10.934
2	Het[A*D*] 1111-1100	120.041	131.782	-11.741
4	Het[AD*] 0000-1100	4.550	14.795	-10.245
5	Het[A*D] 1111-0011	116.310	128.067	-11.756
7	Het[AF] 0000-0101	12.099	24.707	-12.607
8	Het[A*F*] 1111-1010	107.052	120.058	-13.006
10	Het[AF*] 0000-1010	7.583	19.535	-11.952
11	Het[A*F] 1111-0101	107.051	120.052	-13.001
13	Het[AG] 0000-0110	-4.379	7.503	-11.882
14	Het[A*G*] 1111-1001	77.844	89.439	-11.596
16	Het[AG*] 0000-1001	40.894	52.766	-11.872
17	Het[A*G] 1111-0110	144.832	159.096	-14.263
19	Het[DF] 0011-0101	45.968	58.936	-12.968
20	Het[D*F*] 1100-1010	46.770	59.550	-12.780
22	Het[DF*] 0011-1010	34.741	49.654	-14.913
23	Het[D*F] 1100-0101	34.334	49.333	-14.999
25	Het[DG] 0011-0110	40.431	54.283	-13.851
26	Het[D*G*] 1100-1001	37.624	51.067	-13.443
28	Het[DG*] 0011-1001	38.960	52.389	-13.429
29	Het[D*G] 1100-0110	41.572	55.233	-13.661
31	Het[FG] 0101-0110	52.187	67.526	-15.339
32	Het[F*G*] 1010-1001	29.528	44.699	-15.170
34	Het[FG*] 0101-1001	29.528	44.707	-15.179
35	Het[F*G] 1010-0110	52.188	67.526	-15.339
37	Het[BC] 0001-0010	-6.665	6.397	-13.062
38	Het[B*C*] 1110-1101	117.027	130.834	-13.807
40	Het[BC*] 0001-1101	7.644	20.923	-13.279
41	Het[B*C] 1110-0010	35.689	49.205	-13.516
43	Het[BE] 0001-0100	-51.890	-38.198	-13.693
44	Het[B*E*] 1110-1011	86.288	99.906	-13.618
46	Het[BE*] 0001-1011	62.871	75.739	-12.869
47	Het[B*E] 1110-0100	36.108	49.256	-13.148
49	Het[BH] 0001-0111	4.094	17.037	-12.943
50	Het[B*H*] 1110-1000	7.675	20.446	-12.771
52	Het[BH*] 0001-1000	12.739	26.861	-14.122
53	Het[B*H] 1110-0111	110.285	124.791	-14.506
55	Het[CE] 0010-0100	-43.142	-30.534	-12.608
56	Het[C*E*] 1101-1011	89.125	101.758	-12.633
58	Het[CE*] 0010-1011	65.087	79.171	-14.084
59	Het[C*E] 1101-0100	39.481	53.547	-14.066
61	Het[CH] 0010-0111	60.133	73.455	-13.322
62	Het[C*H*] 1101-1000	56.763	69.754	-12.991
64	Het[CH*] 0010-1000	-23.231	-9.642	-13.589
65	Het[C*H] 1101-0111	86.287	102.518	-16.231
67	Het[EH] 0100-0111	11.194	24.678	-13.484
68	Het[E*H*] 1011-1000	8.771	22.085	-13.314
70	Het[EH*] 0100-1000	-35.619	-23.125	-12.494
71	Het[E*H] 1011-0111	117.075	130.891	-13.816

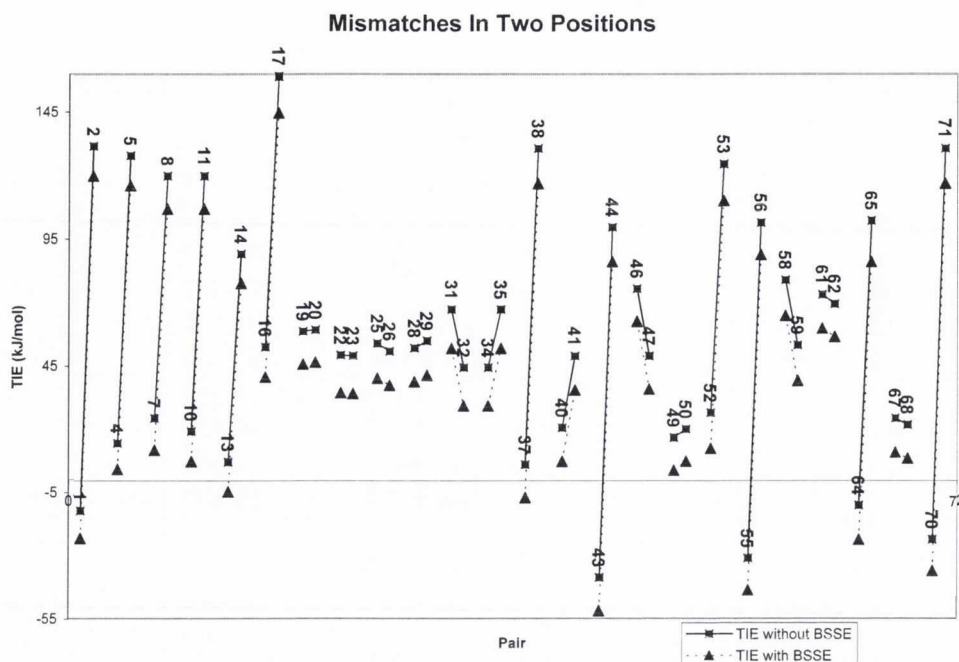


Figure 6.4 TIEs for pairs with two mismatches, with and without BSSE.

For number key see Table 6.4

The TIE of each pair becomes -13.331 kJ/mol more repulsive on average when BSSE is excluded. Whilst the removal of BSSE does alter individual association energies and even changes the energy calculated for two pairs, Het[AG] and Het[BC], from weakly binding (negative) to weakly repulsive (attractive), its effects are not enough to cause any other associations to be considered potentially viable. As seen in section 5.2.2 only three pairs remain in which all interactions (other than those that are complementary) are repulsive. Het[DD*], Het[FF*], Het[GG*]. On the exclusion of BSSE all non-complementary associations are highly repulsive.

6.3.4 BSSE mismatch in four positions

To complete the study on the effect of BSSE, calculations were also performed on pairs with mismatches in all four positions. The results which can be seen below (Table 6.5)(Fig. 6.5) agree with those for the other mismatch sets, BSSE removal shifts the energy values but again does not change the overall energies in relation to one another.

Table 6.5 TIEs for pairs with four mismatches, with and without BSSE

Pair	TIE(kJ/mol)	TIE(kJ/mol) [No BSSE]	BSSE
Het[AA] 0000-0000	86.756	94.623	-7.867
Het[BB] 0001-0001	151.571	163.003	-11.432
Het[CC] 0010-0010	134.624	146.907	-12.283
Het[DD] 0011-0011	223.871	234.714	-10.842
Het[EE] 0100-0100	140.055	152.151	-12.096
Het[FF] 0101-0101	182.290	196.386	-14.096
Het[GG] 0110-0110	217.341	231.550	-14.209
Het[HH] 0111-0111	285.203	297.758	-12.555
Het[A*A*] 1111-1111	385.747	395.652	-9.905
Het[B*B*] 1110-1110	285.200	297.768	-12.568
Het[C*C*] 1101-1101	258.720	270.638	-11.918
Het[D*D*] 1100-1100	224.297	234.878	-10.581
Het[E*E*] 1011-1011	258.720	270.638	-11.918
Het[F*F*] 1010-1010	182.290	196.386	-14.096
Het[G*G*] 1001-1001	190.479	203.717	-13.238
Het[H*H*] 1000-1000	149.867	161.016	-11.149

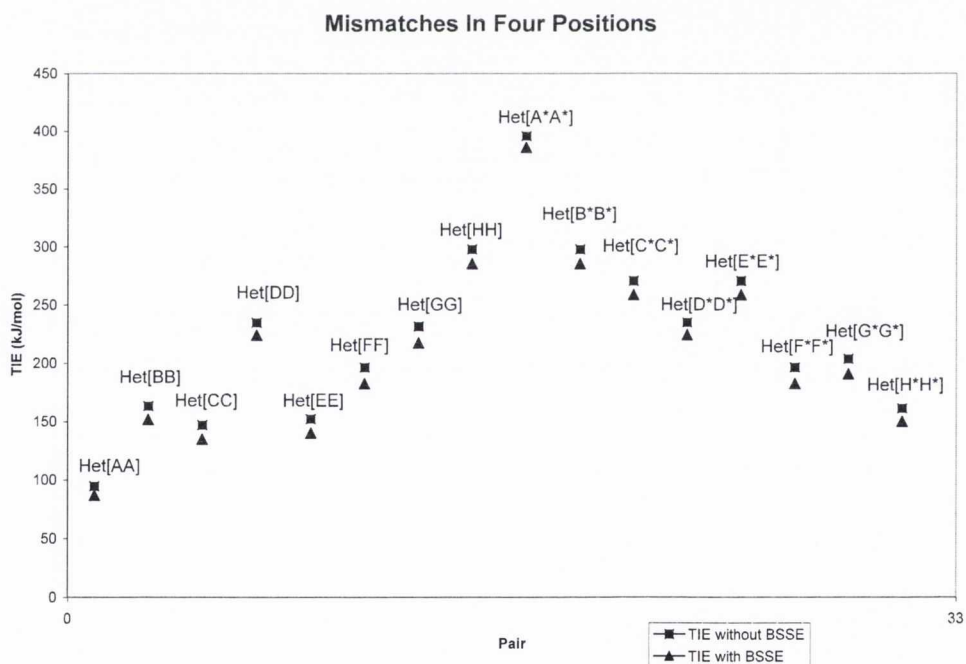


Figure 6.5 TIEs for pairs with four mismatches, with and without BSSE

6.4 Discussion and Summary-BSSE

If a plot is made of all possible interactions for the Het set of letters with and without BSSE (Fig. 6.6) it can be seen that the removal of BSSE from the calculations does not change the overall pattern of results seen. The individual total energy values are shifted to higher (more repulsive) energies but appear in the same sequence.

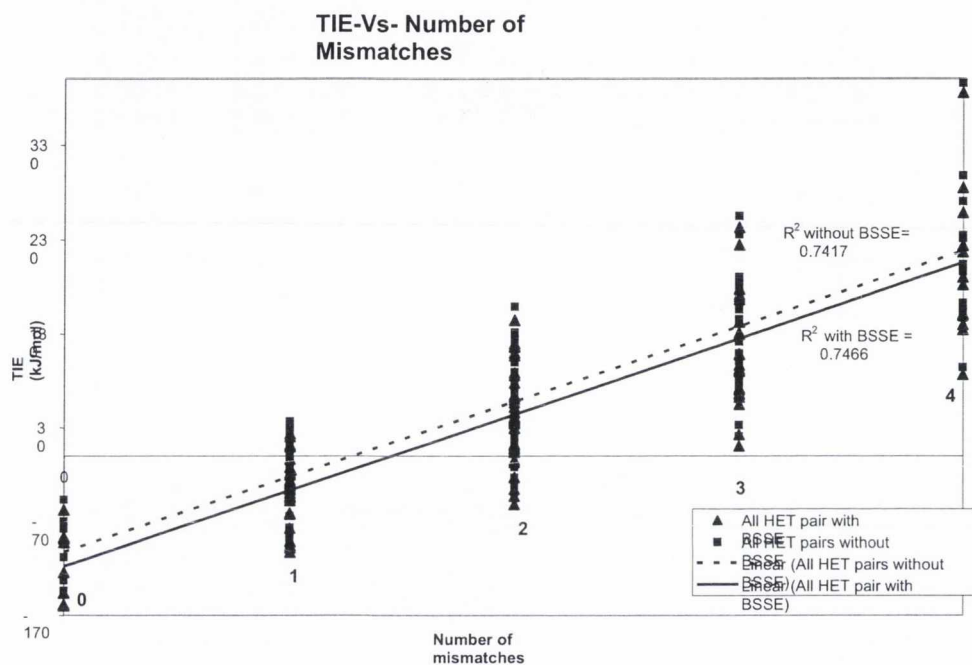


Figure 6.6 Plot of TIE versus number of mismatches, with and without BSSE

If as in section 5.3 only the limiting interactions are considered the same picture is evident with and without the exclusion of BSSE (Fig. 6.7).

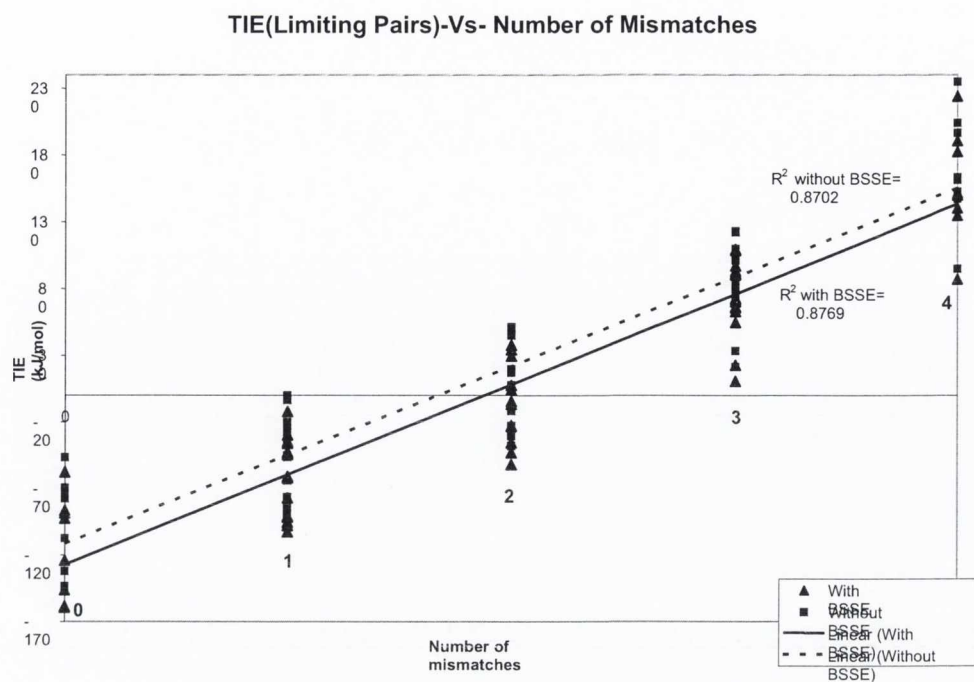


Figure 6.7 Plot of TIE versus number of mismatches in the case of limiting pairs, with and without BSSE

The average BSSE across the entire set of heteronaphthalene associations is 13.218 kJ/mol and has a standard deviation of 1.859 kJ/mol (see appendix A6 for equation). This small deviation indicates that BSSE is almost constant across the entire set of Het letters (Fig. 6.8). The conclusion can be drawn that BSSE does not change the overall sequence of results; its removal simply shifts each pair to a higher energy. As this in the case in further studies of potential alphabets in this thesis BSSE will not systematically be removed from results. This is only possible as an overall pattern and positioning of energies relative to each other is what we seek to determine in this work. BSSE would of course need to be removed if individual accurate energy values are required.

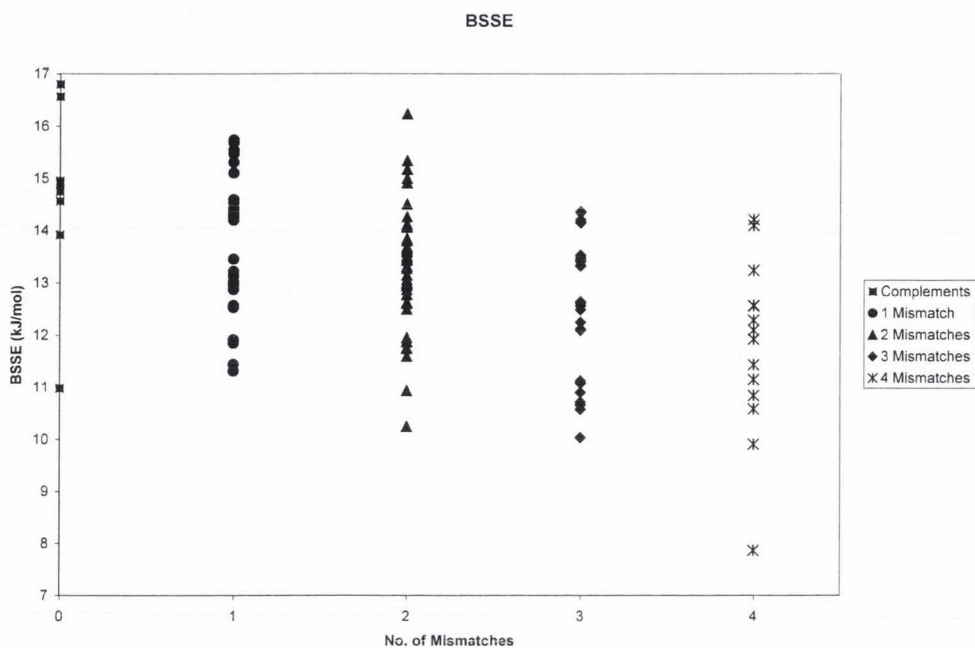


Figure 6.8 BSSE values for all Het pairs grouped by the number of mismatches present.

6.5 Semi-empirical methods

6.5.1 Semi-empirical methods-Introduction

Calculations were performed for all possible Het associations using AM1 and PM3 (See section 2.5) with STRD molecular geometry restrictions in place throughout.

6.6 Semi-empirical Results

6.6.1 Complementary associations

A plot can be made of the TIEs for complementary associations for the two semi-empirical methods (AM1, PM3) and for HF (BSSE removed) (Fig. 6.9).

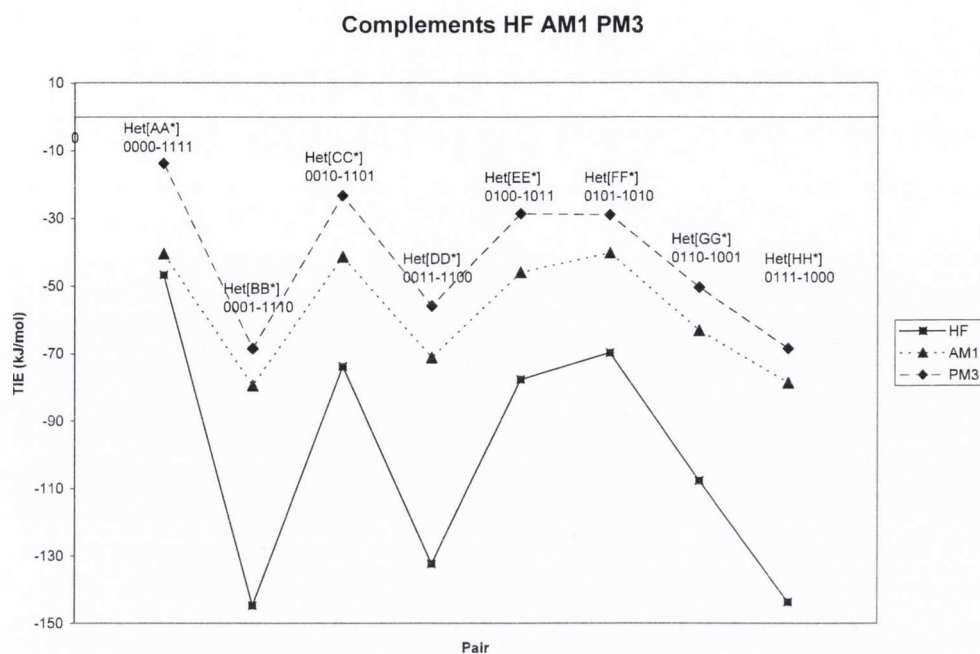


Figure 6.9 TIE Het complementary pairs AM1, PM3 and HF 6-31G* (CP corrected)

The results show that both of the semi-empirical methods give higher energies (more repulsive) than HF (with BSSE removed). PM3 is the most repulsive (most positive) whilst AM1 is positioned in between HF and PM3. Although the individual energy values differ between the three methods all show the same pattern of energies relative to each other, indicating that the HF results give a good representation of the interactions of complementary Het pairs.

6.6.2 Mismatches in one and three positions

In order to use semi-empirical methods to possibly gain some insight into potential consequences of the HF method choice all possible Het interactions need to be considered with both AM1 and PM3.

An overall results plot can be made for pairs that mismatch in one position showing the outcome of the three calculation methods (Fig.6.10).

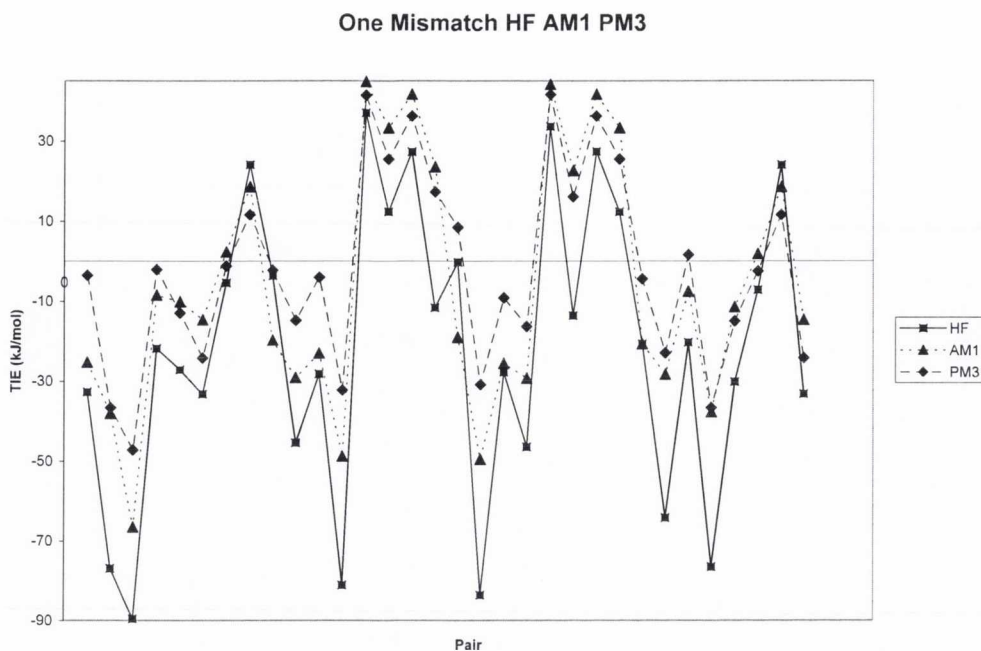


Figure 6.10 TIE Het mismatches in one position AM1, PM3 and HF 6-31G* (CP corrected)

It can be seen that broadly speaking the three methods have roughly the same overall trend in results. For the majority of pairs HF gives the lowest and most binding energies, AM1 and PM3 both give higher energies (agreeing with the results of complementary associations). Some deviation is noted as to which of the semi-empirical methods shows the highest TIE trend. To explore this further the 32 pairs that mismatch in one position can be broken into four groups of eight based on the type and position of the mismatch present. More detailed plots can be made each showing two of these 4 pair subgroups (Fig. 6.11a)(Fig. 6.11b)(Fig. 6.11c)(Fig. 6.11d)(Table 6.6).

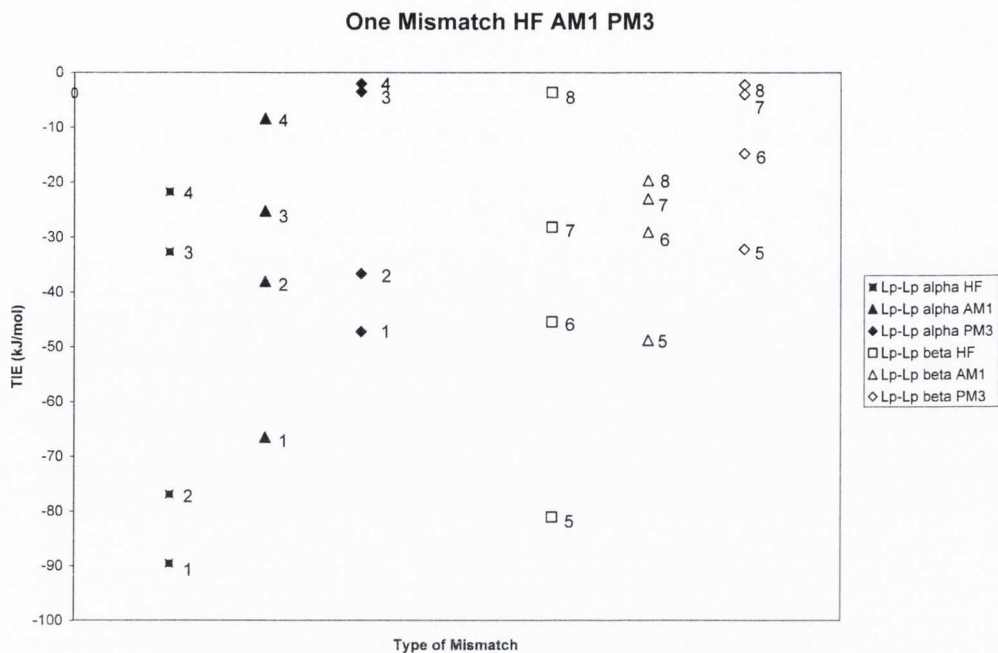


Figure 6.11a TIE Het mismatches in one position AM1, PM3 and HF Lp-Lp mismatch alpha and beta positions. Number Key (Table 6.6)

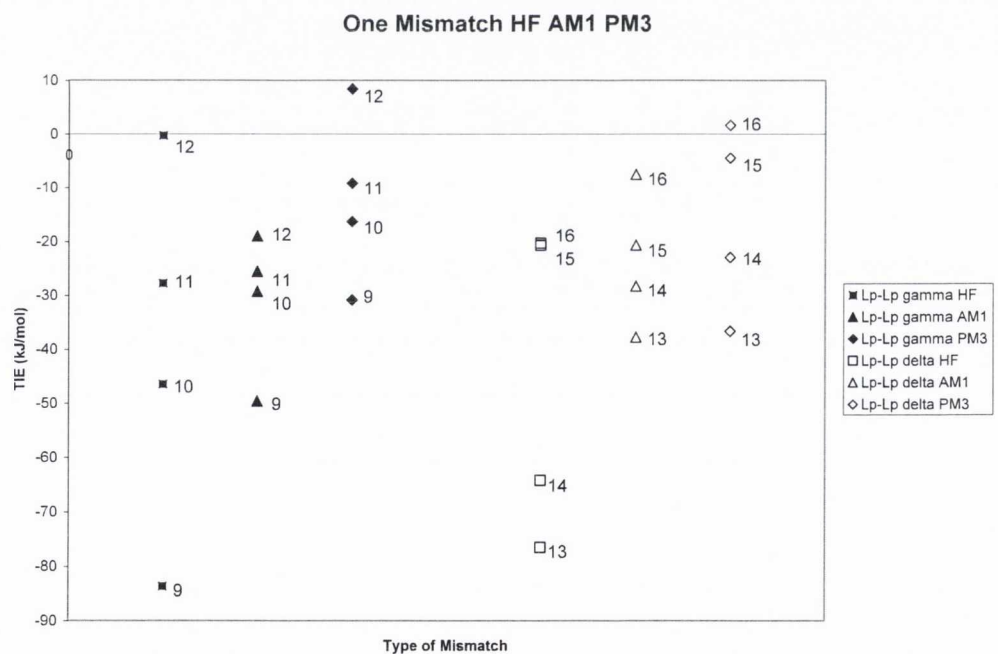


Figure 6.11b TIE Het mismatches in one position AM1, PM3 and HF Lp-Lp mismatch gamma and delta positions. Number Key (Table 6.6)

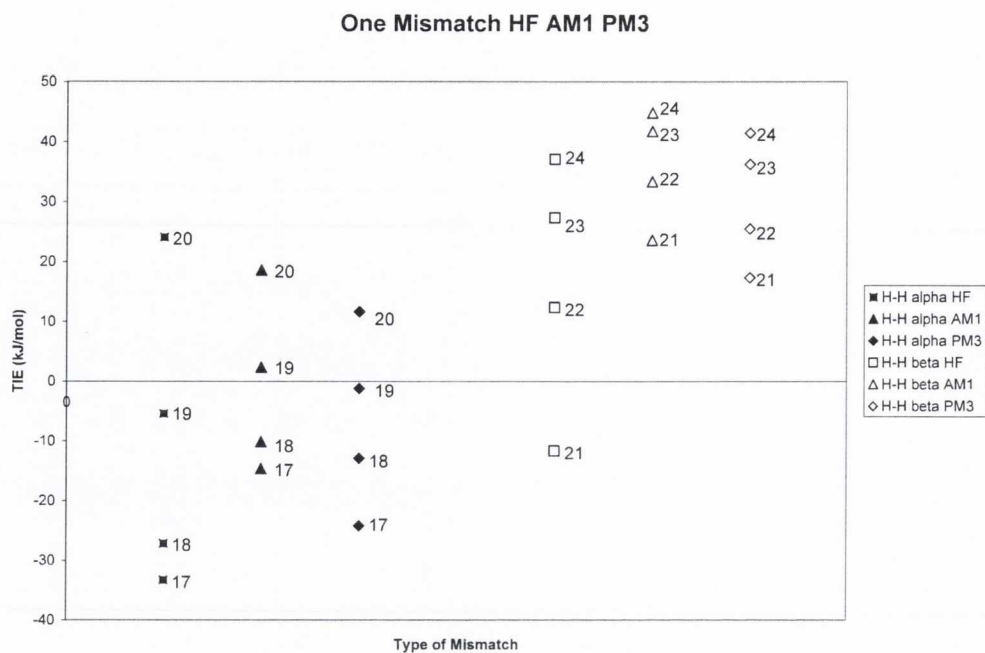


Figure 6.11c TIE Het mismatches in one position AM1, PM3 and HF H-H mismatch alpha and beta positions. Number Key (Table. 6.6)

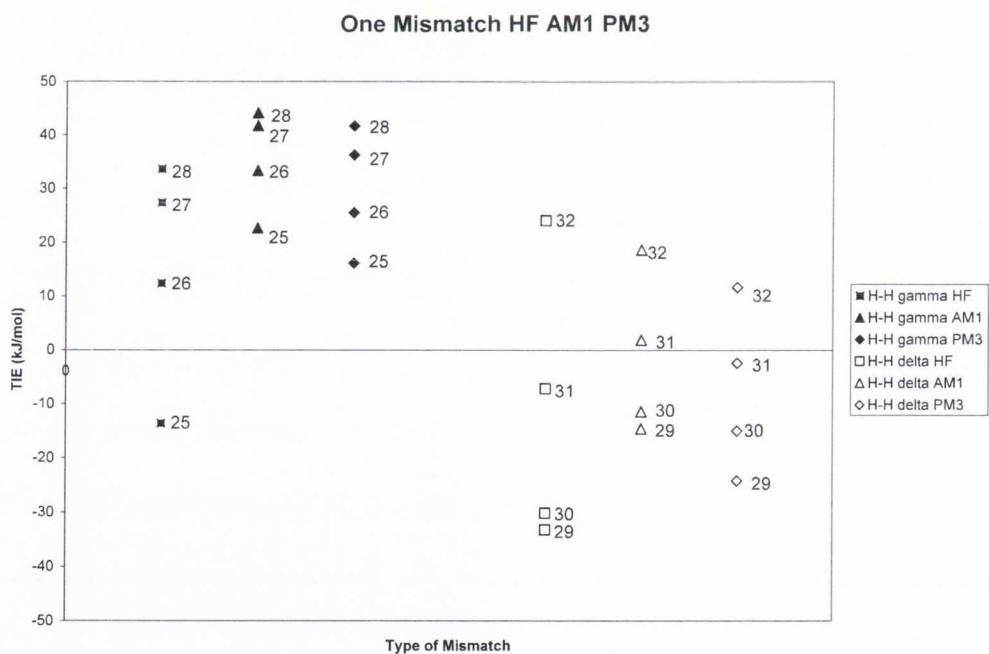


Figure 6.11d TIE Het mismatches in one position AM1, PM3 and HF H-H mismatch gamma and delta positions. Number Key (Table. 6.6)

Table 6.6 Graph number Key. Pairs are ordered from the most binding to most repulsive according to the HF results

	Lp-Lp α	H-H α	
1	Het[DE] 0011-0100	Het[B*G*] 1110-1001	17
2	Het[BG] 0001-0110	Het[A*H*] 1111-1000	18
3	Het[AH] 0000-0111	Het[D*E*] 1100-1011	19
4	Het[CF] 0010-0101	Het[C*F*] 1101-1010	20
	Lp-Lp β	H-H β	
5	Het[DH*] 0011-1000	Het[D*H] 1100-0111	21
6	Het[BF*] 0001-1010	Het[B*F] 1110-0101	22
7	Het[CG*] 0010-1001	Het[C*G] 1101-0110	23
8	Het[AE*] 0000-1011	Het[A*E] 1111-0100	24
	Lp-Lp γ	H-H γ	
9	Het[BD*] 0001-1100	Het[F*H] 1010-0111	25
10	Het[FH*] 0101-1000	Het[B*D] 1110-0011	26
11	Het[EG*] 0100-1001	Het[E*G] 1011-0110	27
12	Het[AC*] 0000-1101	Het[A*C] 1111-0010	28
	Lp-Lp δ	H-H δ	
13	Het[GH*] 0110-1000	Het[G*H] 1001-0111	29
14	Het[CD*] 0010-1100	Het[A*B] 1111-0001	30
15	Het[AB*] 0000-1110	Het[E*F] 1011-0101	31
16	Het[EF*] 0100-1010	Het[C*D] 1101-0011	32

In all three calculation methods H-H repulsion gives higher interaction energies than Lp-Lp. H-H repulsions are strongest when they occur in the middle two positions; this is to be expected as the use of STRD conditions restricts these positions. In general out of the three methods PM3 gives the highest (most repulsive) Lp-Lp values, whilst AM1 values are the highest where H-H repulsions occur. With HF the limiting mismatches were Lp-Lp as all of which were still binding in energy. Two pairs Het[AC*] 0000-1101 and Het[EF*] 0100-1010 each with a Lp-Lp mismatch are shown as weakly repulsive with PM3. This is most likely to be a consequence of PM3 overestimating the Lp-Lp mismatch strength (relative to AM1 and HF). In any case both of these associations have a TIE of < 10 kJ/mol which would most likely not be strong enough to prevent a non-complementary association from forming.

A plot of all three methods being considered in this section can be made for associations that mismatch in three out of a possible four positions (Fig. 6.12).

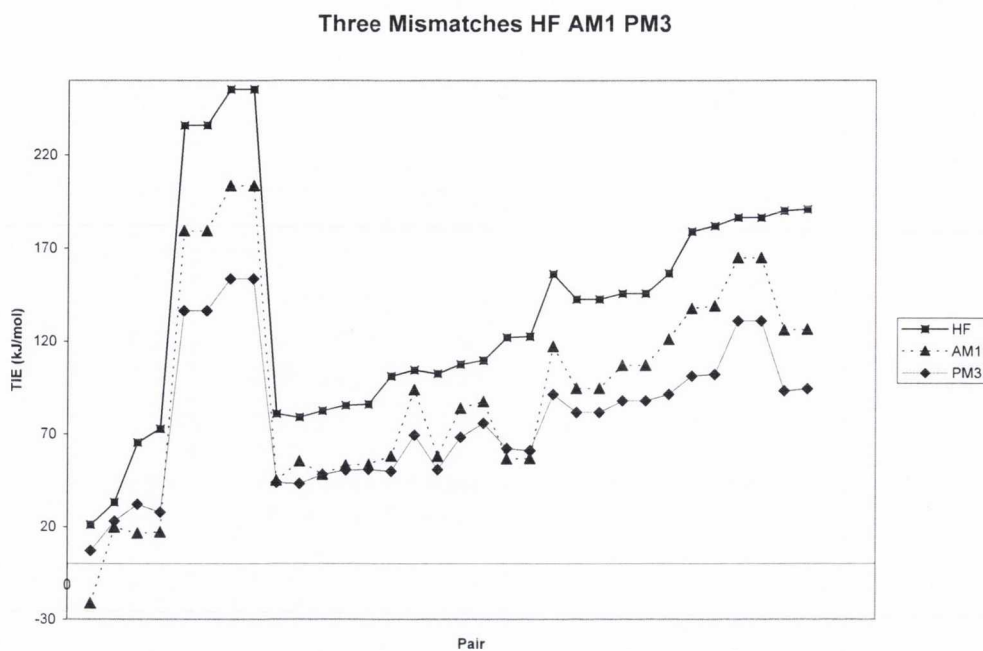


Figure 6.12 TIE Het mismatches in three positions AM1, PM3 and HF 6-31G* (CP corrected)

The results depict a very similar pattern over the three calculation methods. In all cases HF gives the most repulsive interactions (the reverse of what was seen for complementary associations and those that mismatch in one position).

Breaking the association into groups based on the type of mismatch present and examining (in the first instance) associations containing three Lp-Lp mismatches (Fig. 6.13a), Het[AB] 0000-0001, Het[AC] 0000-0010, Het[AE] 0000-0100 and Het[AH*] 0000-1000, AM1 in all cases gives the most attractive TIE and HF the least. The results for the four pairs with three H-H mismatches (Fig. 6.5a), Het[A*B*] 1111,1110, Het[A*C*] 1111-1101, Het[A*E*] 1111,1011 and Het[A*H] 1111-0111, show that HF is still the most repulsive but the ordering of the semi-empirical is now reversed, PM3 shows the least repulsive energies. All other associations that mismatch in three positions contain a mixture of Lp-Lp and H-H mismatches in a 2:1 ratio of one kind of mismatch to another. For these associations HF remains the most repulsive of the three methods, AM1 and PM3 are close to each other in energy and occasionally vary in terms of which method gives the least binding energy (Fig. 6.13b). Only one pair Het[AE] AM1 has weakly binding TIE, this pair has three Lp-Lp interactions. This is most likely due to a Lp-Lp repulsion underestimation relative to PM3 also noted for mismatches in one position.

Table 6.7 Graph number Key. Pairs are ordered from the most binding to most repulsive according to the HF results

1	Lp-Lp Lp-Lp Lp-Lp Het[AE] 0000-0100		
2	Het[AC] 0000-0010		
3	Het[AB] 0000-0001		
4	Het[AH*] 0000-1000		
5	H-H H-H H-H Het[A*C*] 1111-1101		
6	Het[A*E*] 1111-1011		
7	Het[A*B*] 1111-1110		
8	Het[A*H] 1111-0111		
9	Lp-Lp Lp-Lp H-H Het[EF] 0100-0101	21	H-H H-H Lp-Lp Het[B*F*] 1110-1010
10	Het[EG] 0100-0110	22	Het[C*G*] 1101-1001
11	Het[D*E] 1100-0100	23	Het[E*G*] 1011-1001
12	Het[BF] 0001-0101	24	Het[E*F*] 1011-1010
13	Het[F*H*] 1010-1000	25	Het[C*F] 1101-0101
14	Het[D*H*] 1100-1000	26	Het[FH] 0101-0111
15	Het[CG] 0010-0110	27	Het[DH] 0011-0111
16	Het[BD] 0001-0011	28	Het[B*D*] 1110-1100
17	Het[CF*] 0010-1010	29	Het[B*G] 1110-0110
18	Het[CD] 0010-0011	30	Het[GH] 0110-0111
19	Het[G*H*] 1001-1000	31	Het[DE*] 0011-1011
20	Het[BG*] 0001-1001	32	Het[C*D*] 1101-1100

Although on comparison of the results for the three computation methods HF, AM1 and PM3, do show a wide variation in calculated absolute energy values, the overall pattern of relative energies remains very similar. The absolute picture remains unchanged. A Het alphabet that contains the possibility of mismatches in one and three positions is not energetically viable.

6.6.3 Mismatches in two positions

On comparing the results of the three methods for pairs which mismatch in two positions (Fig. 6.14) a similar overall pattern can be identified, however at many points it is somewhat unclear exactly what is happening to the results pattern seen for each method.

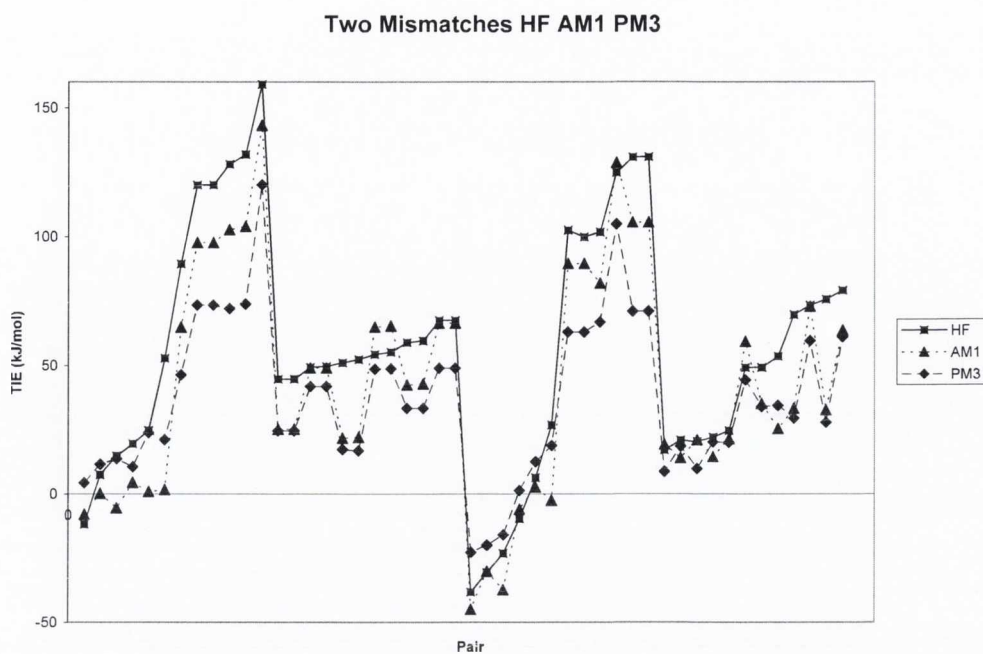


Figure 6.14 TIE Het mismatches in two positions AM1, PM3 and HF 6-31G* (CP corrected)

In order to gain a clearer picture two plots were made breaking the results into groups based on parity and mismatch type (Fig. 6.15a) (Fig. 6.15b) (Table 6.8).

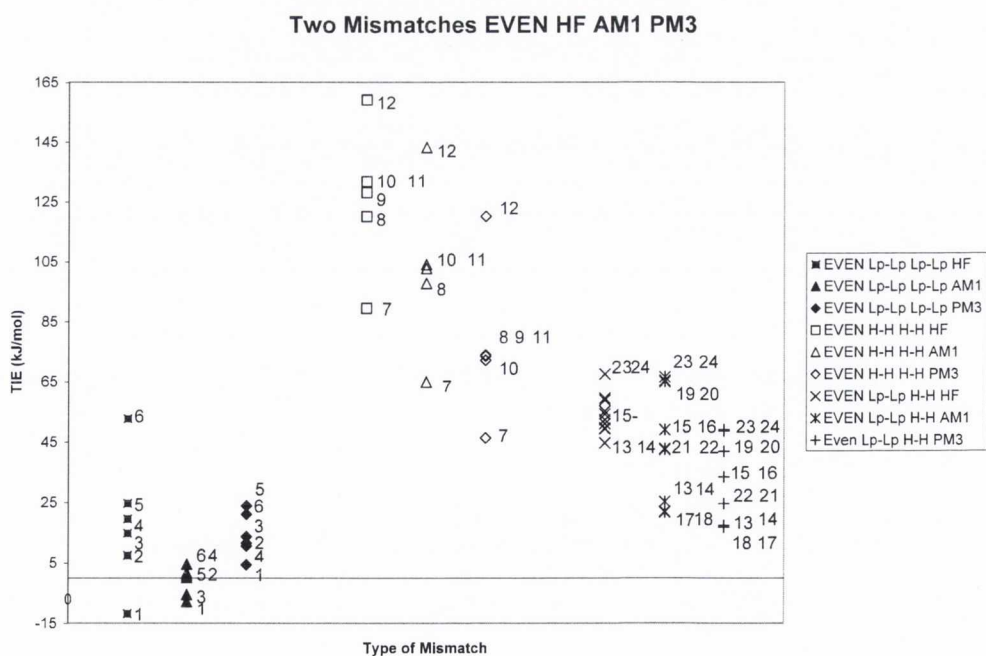


Figure 6.15a TIE Het even parity mismatches in two positions AM1, PM3 and HF Number Key (Table 6.8)

Two Mismatches ODD HF AM1 PM3

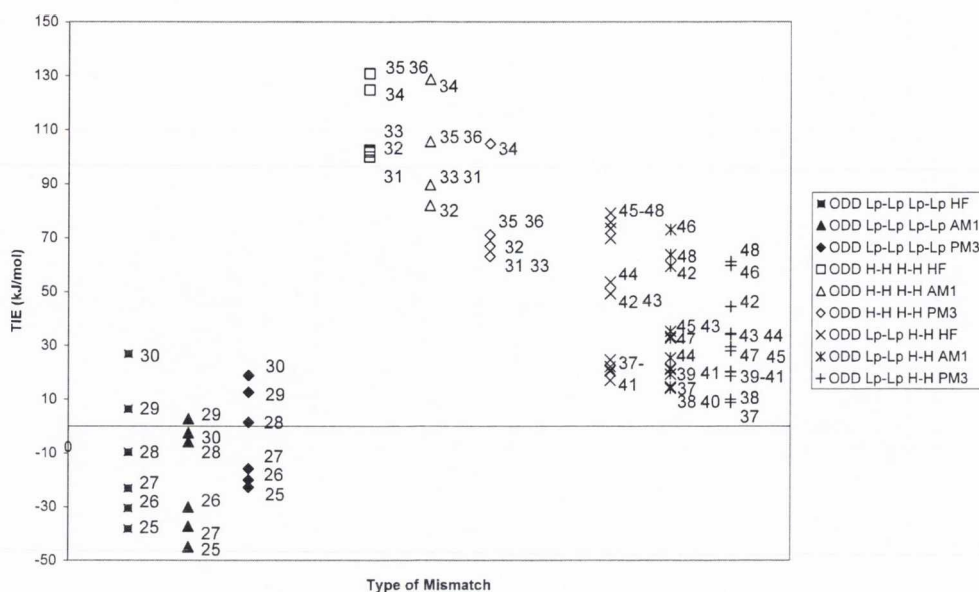


Figure 6.15b TIE Het odd parity mismatches in two positions AM1, PM3 and HF Number Key (Table 6.8)

Table 6.8 Graph number Key. Pairs are ordered from the most binding to most repulsive according to the HF results

1	Even Lp-Lp Lp-Lp Het[AD] 0000-0011	25	Odd Lp-Lp Lp-Lp Het[BE] 0001-0100
2	Het[AG] 0000-0110	26	Het[CE] 0010-0100
3	Het[AD*] 0000-1100	27	Het[EH*] 0100-1000
4	Het[AF*] 0000-1010	28	Het[CH*] 0010-1000
5	Het[AF] 0000-0101	29	Het[BC] 0001-0010
6	Het[AG*] 0000-1001	30	Het[BH*] 0001-1000
7	Even H-H H-H Het[A*G*] 1111-1001	31	Odd H-H H-H Het[C*H] 1101-0111
8	Het[A*F] 1111-0101	32	Het[B*E*] 1110-1011
9	Het[A*F*] 1111-1010	33	Het[C*E*] 1101-1011
10	Het[A*D] 1111-0011	34	Het[B*H] 1110-0111
11	Het[A*D*] 1111-1100	35	Het[B*C*] 1110-1101
12	Het[A*G] 1111-0110	36	Het[E*H] 1011-0111
13	Even Lp-Lp H-H Het[F*G*] 1010-1001	37	Odd Lp-Lp H-H Het[BH] 0001-0111
14	Het[FG*] 0101-1001	38	Het[BC*] 0001-1101
15	Het[D*F] 1100-0101	39	Het[B*H*] 1110-1000
16	Het[DF*] 0011-1010	40	Het[E*H*] 1011-1000
17	Het[D*G*] 1100-1001	41	Het[EH] 0100-0111
18	Het[DG*] 0011-1001	42	Het[B*C] 1110-0010
19	Het[DG] 0011-0110	43	Het[B*E] 1110-0100
20	Het[D*G] 1100-0110	44	Het[C*E] 1101-0100
21	Het[DF] 0011-0101	45	Het[C*H*] 1101-1000
22	Het[D*F*] 1100-1010	46	Het[CH] 0010-0111
23	Het[FG] 0101-0110	47	Het[BE*] 0001-1011
24	Het[F*G] 1010-0110	48	Het[CE*] 0010-1011

Examining the results in terms of mismatch type (for even and odd parities) reveals that for Lp-Lp mismatches AM1 (in the majority of cases) gives the lowest of the three calculation methods, whilst in the case of H-H mismatches the same is true of PM3. This relative ordering of methods mirror those seen for mismatches in three positions. It appears that AM1 underestimates Lp-Lp mismatches (or overestimates H-H) relative to PM3, whilst PM3 underestimates H-H (or overestimates Lp-Lp) relative to AM1. For associations with both a Lp-Lp and a H-H mismatch a more even result can be seen across the three calculation methods.

The HF results indicated that six letters Het[D], Het[D*], Het[F], Het[F*], Het[G], Het[G*] could possibly coexist as every interaction between them other than complementary associations is strongly repulsive. This result is confirmed by both of the semi-empirical methods (Table. 6.9).

Table 6.9 All possible pairing between Het[D], Het[D*],Het[F],Het[F*],Het[G],Het[G*]

Pair	HF CP TIE (kJ/mol)	AM1 TIE (kJ/mol)	PM3 TIE(kJ/mol)
Het[DF] 0011-0101	58.936	42.489	33.418
Het[D*F*] 1100-1010	59.550	42.827	41.894
Het[DF*] 0011-1010	49.654	49.112	33.397
Het[D*F] 1100-0101	49.333	49.057	41.894
Het[DG] 0011-0110	54.283	64.965	48.643
Het[D*G*] 1100-1001	51.067	21.769	16.815
Het[DG*] 0011-1001	52.389	22.075	17.397
Het[D*G] 1100-0110	55.233	65.392	48.756
Het[FG] 0101-0110	67.526	66.488	49.066
Het[F*G*] 1010-1001	44.699	25.234	24.614
Het[FG*] 0101-1001	44.707	25.230	24.614
Het[F*G] 1010-0110	67.526	66.488	49.070

A larger subgroup was found for PM3 composed of Het[A], Het[A*], Het[D], Het[D*], Het[F], Het[F*], Het[G], Het[G*] suggesting that 8 letters could possibly coexist (all even parity letters)(Table 6.10).

Table 6.10 All possible pairing between Het[A], Het[A*], Het[D], Het[D*],Het[F],Het[F*],Het[G],Het[G*]

Pair	PM3 TIE(kJ/mol)	Pair	PM3 TIE(kJ/mol)
Het[AD] 0000-0011	4.397	Het[DF] 0011-0101	33.418
Het[AD*] 0000-1100	13.724	Het[DF*] 0011-1010	41.894
Het[A*D*] 1111-1100	73.885	Het[D*F*] 1100-1010	33.397
Het[A*D] 1111-0011	72.141	Het[DF*] 0011-1010	41.894
Het[AF] 0000-0101	23.907	Het[DG] 0011-0110	48.643
Het[AF*] 0000-1010	10.623	Het[DG*] 0011-1001	16.815
Het[A*F*] 1111-1010	73.521	Het[D*G*] 1100-1001	17.397
Het[A*F] 1111-0101	73.521	Het[D*G] 1100-0110	48.756
Het[AG] 0000-0110	11.569	Het[FG] 0101-0110	49.066
Het[AG*] 0000-1001	21.121	Het[FG*] 0101-1001	24.614
Het[A*G*] 1111-1001	46.396	Het[F*G*] 1010-1001	24.614
Het[A*G] 1111-0110	120.110	Het[F*G] 1010-0110	49.070

The limiting association in this larger subgroup is Het[AD] which has only weakly repulsive TIE. Although technically this larger subgroup meets the requirement that all interactions other than those that are complementary should be repulsive, 4 kJ/mol is only very weakly repulsive. PM3 in the case of Het[AD] overestimates the two Lp-Lp interactions compared to HF and AM1 which both consider this pairing to still be attractive. The same point could be made about a subgroup that emerges again for PM3 but this time for an odd parity set of letters (Table 6.11). The limiting pair Het[CH*] is repulsive but not strongly enough to make this set of letters viable.

Table 6.11 All possible pairing between Het[B], Het[B*], Het[C], Het[C*],Het[H],Het[H*]

Pair	PM3 TIE(kJ/mol)
Het[BC] 0001-0010	12.619
Het[BC*] 0001-1101	18.548
Het[B*C*] 1110-1101	71.099
Het[B*C] 1110-0010	44.401
Het[BH] 0001-0111	8.757
Het[BH*] 0001-1000	18.861
Het[B*H*] 1110-1000	9.945
Het[B*H] 1110-0111	104.838
Het[CH] 0010-0111	59.639
Het[CH*] 0010-1000	1.377
Het[C*H*] 1101-1000	29.493
Het[C*H] 1101-0111	63.032

6.6.4 Mismatches in four positions

To complete the set of method comparison tests calculations were also performed for mismatches in all four positions (Fig. 6.16)(Table 6.12). As the number of H-H mismatches increases, the order of results becomes PM3-AM1-HF (from least to most repulsive) agreeing with what was seen for three H-H mismatches when considering mismatches in three positions. It is noted due to the underestimation of Lp-Lp interactions by AM1 (relative to PM3) and the overestimation of H-H repulsions (relative to PM3) that the order of methods reverses as the number of H-H mismatches increases.

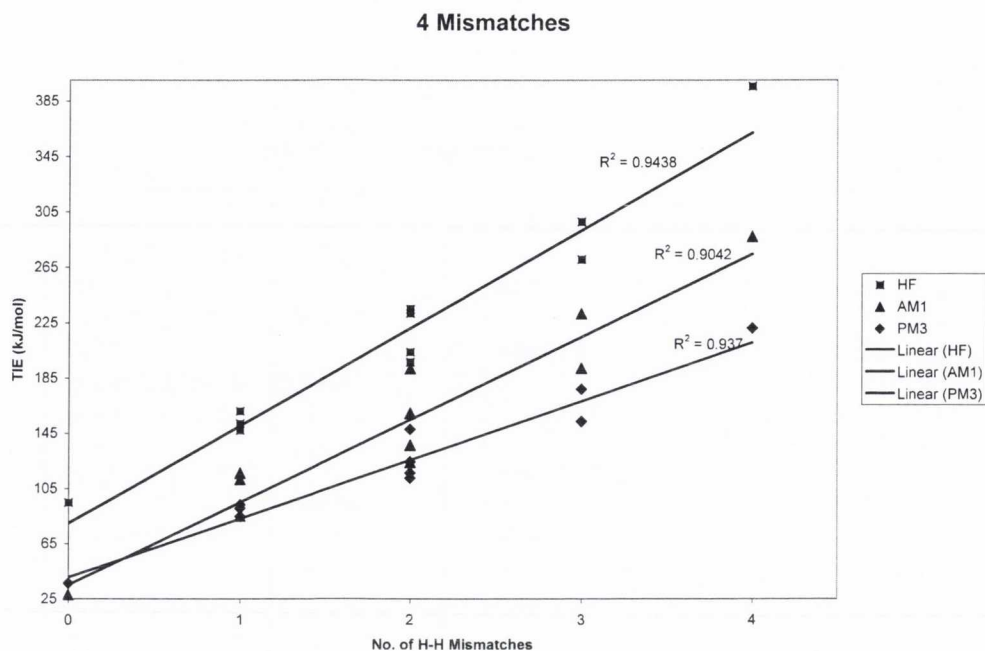


Figure 6.16 TIE Het pairs with four mismatches, AM1, PM3 and HF. Key (Table 6.12)

Table 6.12 Het pairs that mismatch in 4 positions detailing the number of H-H mismatches per pair

AA 0000-0000	0 H-H
BB 0001-0001	1 H-H
CC 0010-0010	1 H-H
EE 0100-0100	1 H-H
H*H* 1000-1000	1 H-H
DD 0011-0011	2 H-H
FF 0101-0101	2 H-H
GG 0110-0110	2 H-H
D*D* 1100-1100	2 H-H
F*F* 1010-1010	2 H-H
G*G* 1001-1001	2 H-H
HH 0111-0111	3 H-H
B*B* 1110-1110	3 H-H
C*C* 1101-1101	3 H-H
E*E* 1011-1011	3 H-H
A*A* 1111-1111	4 H-H

6.7 Discussion and Summary-Semi-empirical results

As in previous Het overall results sections plots can be made of all associations grouped by the number of mismatches versus the TIE for each pair. Both AM1 (Fig. 6.17) and PM3 plots (Fig. 6.18) show that as expected the TIE increases in value as the number of mismatches increases. As with HF a large spread of values is seen at each mismatches grouping. This spread of values indicates that the range of TIE values for a given mismatch

is not unique to that mismatch; a strong overlap between mismatch values in different groups is evident. The R^2 value calculated for the trend line of HF 6-31G* CP results is 0.7466 higher than both of the semi-empirical methods. This indicates a larger spread of results is seen for the results determined using semi-empirical methods. Although using a semi-empirical method to study the Het potential alphabet does result in a change in the absolute TIE determined, importantly it does not change the overall outcome.

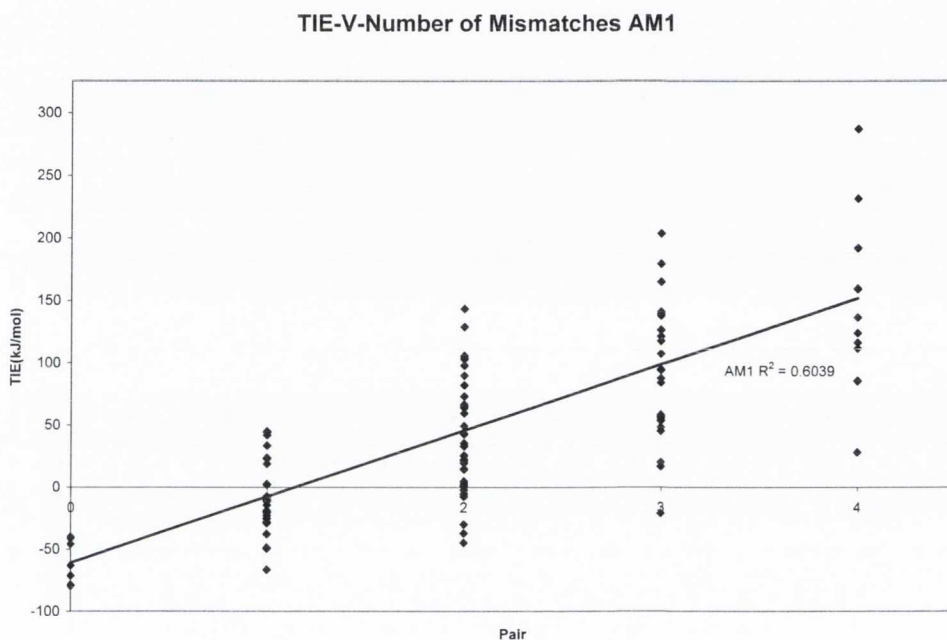


Figure 6.17 Graph showing total interaction energy against number of mismatches for all pairs AM1.

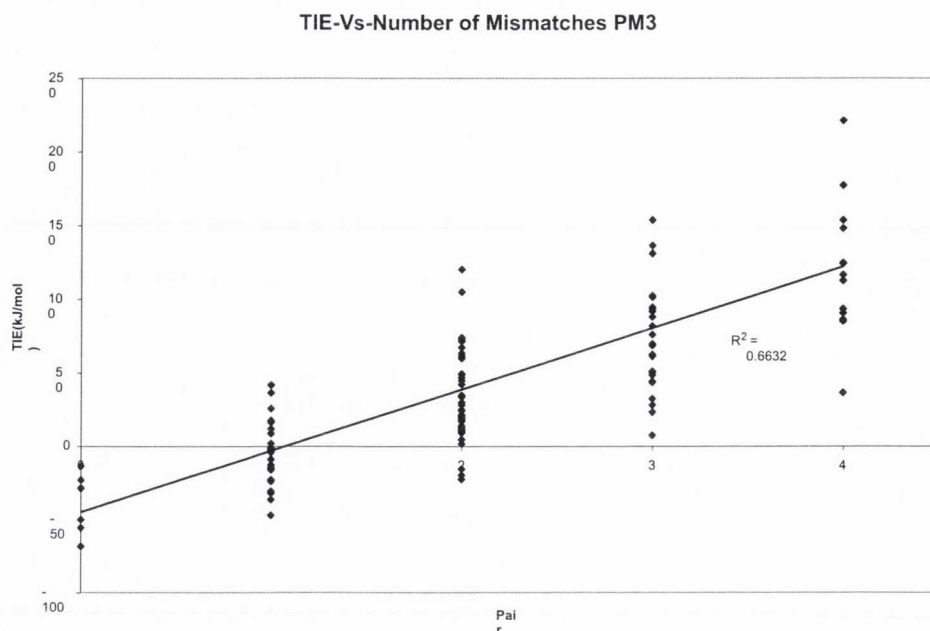


Figure 6.18 Graph showing total interaction energy against number of mismatches for all pairs PM3.

A simplification to the data set can be made by considering only the limiting pairs (see section 5.3) for each method. Once again both methods AM1 (Fig. 6.19) and PM3 (Fig. 6.20) show a spread of results that overlaps into the range of TIE values for a set of associations with a different number of mismatches. Both of the semi-empirical methods show a less linear result than that of HF where $R^2 = 0.8769$.

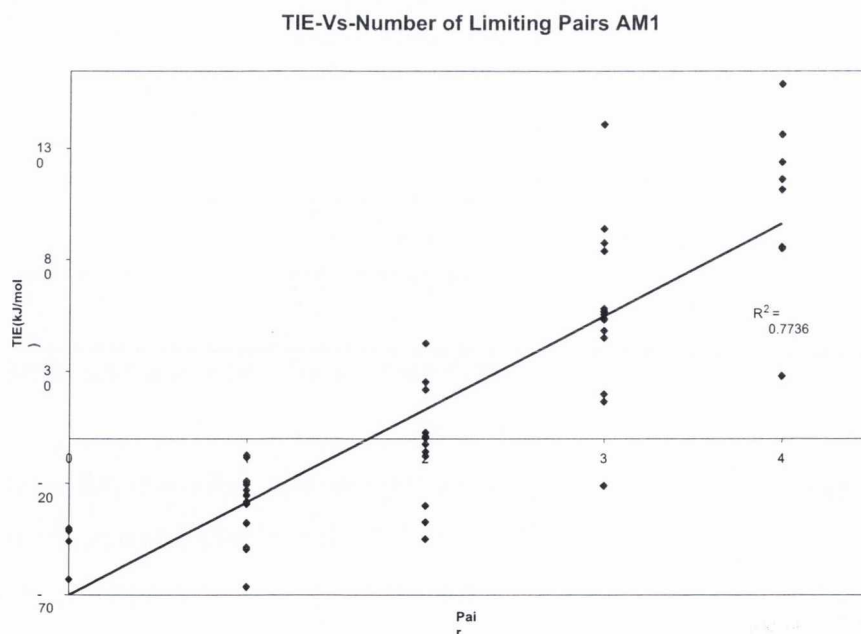


Figure 6.19 Graph showing total interaction energy against number of mismatches for all pairs AM1. Limiting interactions only

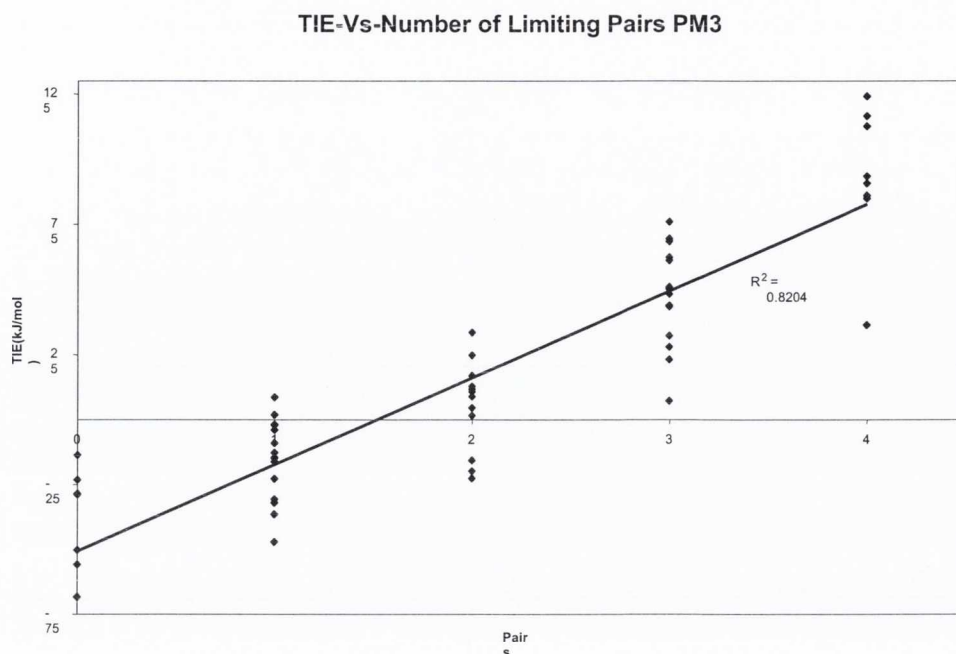


Figure 6.20 Graph showing total interaction energy against number of mismatches for all pairs AM1. Limiting interactions only

6.8 MP2

6.8.1 Introduction

As discussed in section 2.6 electrons do not move entirely independently of each other, they move in a correlated fashion in order to avoid each other. The methods considered so far in this thesis during the exploration of the Het molecules, semi-empirical (AM1, PM3) and Hartree-Fock, do not take correlation into account. To assess the effect of correction on the Het set of letters and to further explore the patterns of relative energies seen, MP2 calculations were carried out and compare to the HF results. STRD geometry constraints were in place throughout.

6.9 MP2 Results

6.9.1 Complementary Pairs

In order to gain some insight into how much correlation affects the results gathered for the set of Het molecules being considered, MP2 calculations were performed and the results compared to those seen for HF (Table 6.13).

Table 6.13 TIE for Het complementary associations. OPT=Optimisation, STRD=Standard conditions, CP=Counterpoise, SP=Single Point, HFG=Hartree-Fock Geometry

Pair	TIE (kJ/mol) HF OPT 6-31G* STRD	TIE (kJ/mol) [H] HF OPT 6-31G* STRD CP	CP HF	Difference [H-M] Correlation
Het[AA*]	-57.718	-46.739	10.979	19.414
Het[BB*]	-159.610	-144.716	14.894	16.234
Het[CC*]	-88.338	-73.778	14.560	21.607
Het[DD*]	-146.232	-132.318	13.914	21.801
Het[EE*]	-92.378	-77.632	14.746	22.729
Het[FF*]	-86.551	-69.749	16.803	25.353
Het[GG*]	-124.236	-107.674	16.562	23.134
Het[HH*]	-158.645	-143.693	14.952	16.698
Pair	TIE (kJ/mol) MP2 OPT 6-31G* STRD	TIE (kJ/mol)[M] MP2 OPT 6-31G* STRD CP	CP MP2	TIE (kJ/mol) MP2 SP HFG 6-31G*
Het[AA*]	-93.007	-66.154	26.854	-93.370
Het[BB*]	-196.840	-160.950	35.890	-197.485
Het[CC*]	-129.924	-95.385	34.540	-129.561
Het[DD*]	-187.674	-154.119	33.555	-188.257
Het[EE*]	-135.006	-100.361	34.645	-135.174
Het[FF*]	-134.195	-95.102	39.093	-133.190
Het[GG*]	-169.202	-130.808	38.395	-168.936
Het[HH*]	-196.066	-160.391	35.676	-196.991

The results determined using MP2 are lower (more attractive) than the corresponding HF results. This shift towards more negative (binding) TIE is to be expected as MP2 includes the correlated movements of electrons neglected by HF causing an overestimation in repulsiveness (section 2.6). The average difference between the methods (MP2 and HF OPT STRD) is 41 kJ/mol. When BSSE is removed this reduces the attractiveness of pairs by more than twice (average of 35 kJ/mol) the amount seen for HF (average of 14kJ/mol). Comparing both methods with BSSE taken into account the average change in TIE due to correlation per complementary pair is 21 kJ/mol. Very little change in energy was seen on comparing the results of MP2 optimisation and MP2 Single Point (SP) at the Hartree-Fock geometry (HFG) calculations, which implies that only a small change in geometry exists between the dimer structures determined by HF and MP2. As this is the case and considering that MP2 calculations are costlier in terms of calculation time (see appendix

A7), further MP2 calculation will be performed as SP at the HFG. A plot can be made comparing the TIEs calculated with MP2 and HF (Fig. 6.21).

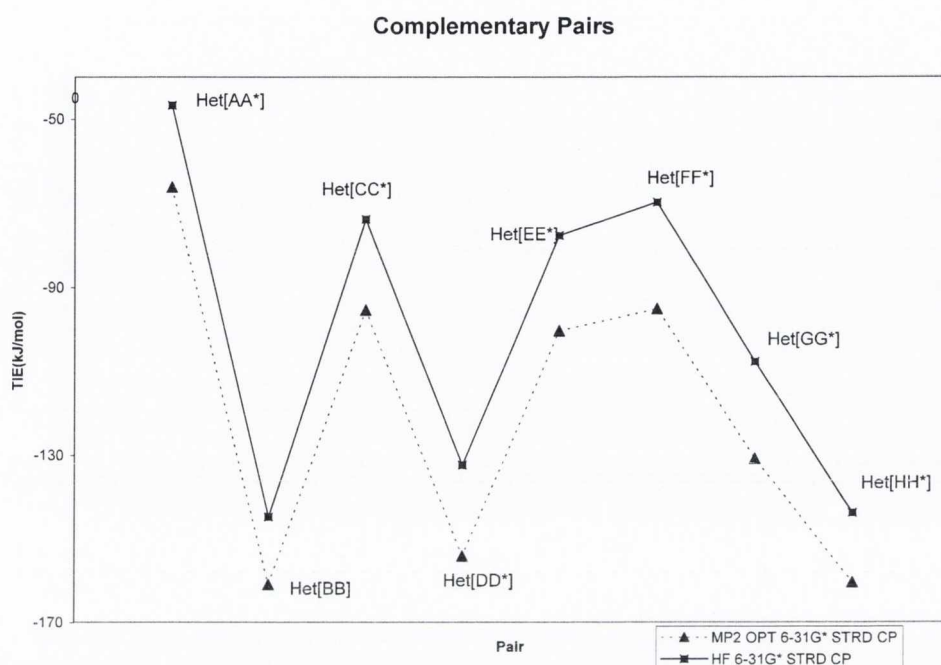


Figure 6.21 TIE for Het complementary associations. MP2 and HF 6-31G* STRD CP.

In order from most repulsive to least.

No change in the overall relative pattern of energies is seen on comparing the calculation methods.

6.8.2 Non- Complementary Pairs

It is important to have confidence that the results collected for all possible associations at the HF level do in fact capture the behaviour of this hypothetical Het set of letters.

Although the results for complementary pairs show agreement (between HF and MP2) in the pattern of results it is crucial to investigate if this is the case for other associations that are non-complementary. In this section a representative sample subset of pairs in each $\bar{\delta}$ group will be used to try and determine the behaviour of mismatching associations with MP2. Each subset has been constructed to include one of each class of mismatch grouped by type and XNOR value. In this study pairs are not attached to a backbone like structure, this makes some hydrogen bonding positions are interchangeable and means that not every possible combination needs to be represented in each subset. A Lp-Lp mismatch, for

example, in the α position is the same as a Lp-Lp mismatch in the δ position as both are terminal free positions. The same simplification can be applied to the β and γ positions as they are both constrained inner constrained positions. This logic only applies in the absence of the molecules being anchored to a backbone like structure. If a backbone structure were present the two inner (or outer) hydrogen bonding positions would not be interchangeable in this way. The pairs chosen for each number of mismatches can be seen in the table below (Table 6.14).

Table 6.14 Subsets of pairs grouped by class and number of mismatches

1 Mismatch	
XNOR 1000 $\alpha(=\delta)$ Lp-Lp	Het[DE] 0011-0100
XNOR 0100 $\beta(=\gamma)$ Lp-Lp	Het[DH*] 0011-1000
XNOR 1000 $\alpha(=\delta)$ H-H	Het[D*E*] 1100-1011
XNOR 0100 $\beta(=\gamma)$ H-H	Het[D*H] 1100-0111
2 Mismatches	
XNOR 1100 $\alpha\beta(=\gamma\delta)$ Lp-Lp Lp-Lp	Het[DA] 0011-0000
XNOR 0110 $\beta\gamma$ Lp-Lp Lp-Lp	Het[G*A] 1001-0000
XNOR 1010 $\alpha\gamma(=\beta\delta)$ Lp-Lp Lp-Lp	Het[BE] 0001-0100
XNOR 1001 $\alpha\delta$ Lp-Lp Lp-Lp	Het[GA] 0110-0000
XNOR 1100 $\alpha\beta(=\gamma\delta)$ H-H H-H	Het[D*A*] 1100-1111
XNOR 0110 $\beta\gamma$ H-H H-H	Het[GA*] 0110-1111
XNOR 1010 $\alpha\gamma(=\beta\delta)$ H-H H-H	Het[B*E*] 1110-1011
XNOR 1001 $\alpha\delta$ H-H H-H	Het[G*A*] 1001-1111
XNOR 1001 $\alpha\beta(=\gamma\delta)$ Lp-Lp H-H	Het[EH] 0100-0111
XNOR 0110 $\beta\gamma$ Lp-Lp H-H	Het[F*D] 1010-0011
XNOR 1010 $\alpha\gamma(=\beta\delta)$ Lp-Lp H-H	Het[HC] 0111-0011
XNOR 1001 $\alpha\delta$ Lp-Lp H-H	Het[FD] 0101-0011
3 Mismatches	
XNOR 1110 $\alpha\beta\gamma(=\beta\gamma\delta)$ Lp-Lp Lp-Lp Lp-Lp	Het[AB] 0000-0001
XNOR 1011 $\alpha\gamma\delta(=\alpha\beta\delta)$ Lp-Lp Lp-Lp Lp-Lp	Het[AE] 0000-0100
XNOR 1110 $\alpha\beta\gamma(=\beta\gamma\delta)$ H-H H-H H-H	Het[A*B*] 1111-1110
XNOR 1011 $\alpha\gamma\delta(=\alpha\beta\delta)$ H-H H-H H-H	Het[A*E*[1111-1011]]
XNOR 1110 $\alpha\beta\gamma(=\beta\gamma\delta)$ Lp-Lp Lp-Lp H-H	Het[CD] 0010-0011
XNOR 0010 $\alpha\gamma\delta(=\alpha\beta\delta)$ Lp-Lp Lp-Lp H-H	Het[BD] 0001-0011
XNOR 0111 $\alpha\beta\gamma(=\beta\gamma\delta)$ Lp-Lp H-H Lp-Lp	Het[CF*] 0010-1010
XNOR 1101 $\alpha\gamma\delta(=\alpha\beta\delta)$ Lp-Lp H-H Lp-Lp	Het[EG] 0100-0110
XNOR 1110 $\alpha\beta\gamma(=\beta\gamma\delta)$ H-H H-H Lp-Lp	Het[C*D*] 1101-1100
XNOR 1101 $\alpha\gamma\delta(=\alpha\beta\delta)$ H-H H-H Lp-Lp	Het[B*D*] 1110-1100
XNOR 0111 $\alpha\beta\gamma(=\beta\gamma\delta)$ H-H Lp-Lp H-H	Het[FC*] 0101-1101
XNOR 1101 $\alpha\gamma\delta(=\alpha\beta\delta)$ H-H Lp-Lp H-H	Het[E*G*] 1011-1001
4 Mismatches XNOR 1111	
$\alpha\beta\gamma\delta$ Lp-Lp Lp-Lp Lp-Lp Lp-Lp	Het[AA] 0000-0000
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ Lp-Lp Lp-Lp Lp-Lp H-H	Het[BB] 0001-0001
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ Lp-Lp Lp-Lp H-H Lp-Lp	Het[CC] 0010-0010
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ Lp-Lp Lp-Lp H-H H-H	Het[DD] 0011-0011
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ Lp-Lp H-H Lp-Lp H-H	Het[FF] 0101-0101
$\alpha\beta\gamma\delta$ Lp-Lp H-H H-H Lp-Lp	Het[GG] 0110-0110
$\alpha\beta\gamma\delta$ H-H H-H H-H H-H	Het[A*A*] 1111-1111
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ H-H H-H H-H Lp-Lp	Het[B*B*] 1110-1110
$\alpha\beta\gamma\delta(\delta\gamma\beta\alpha)$ H-H H-H Lp-Lp H-H	Het[C*C*] 1101-1101
$\alpha\beta\gamma\delta$ H-H Lp-Lp Lp-Lp H-H	Het[G*G*] 1001-1001

Using the subset of pairs set out in the table above, MP2 6-31G* SP calculations were performed for each association at the HFG and CP was added after optimization to remove BSSE. In a few test cases pairs were optimized using MP2 to confirm that very little difference in energy is seen on comparing the SP and optimization results. The results are

shown in Table 6.15a-d. The data shown in this section is for MP2 with a frozen core, the default setting in Gaussian 03W (for discussion and exploration of this see appendix A8).

Table 6.15a Subsets of pairs that mismatch in one position. MP2 and HF

Fig. 6.22 Number key	1 Mismatch	TIE MP2 (kJ/mol)	TIE HF (kJ/mol)
1	Het[D*E*] SP MP2 6-31G*	-65.609	-18.296
	Het[D*E*] CP	-35.672	-5.424
	Het[D*E*] OPT MP2 6-31G*	-66.118	
2	Het[DE] SP MP2 6-31G*	-142.102	-102.683
	Het[DE] CP	-113.407	-89.558
	Het[DE] OPT MP2 6-31G*	-141.237	
3	Het[D*H] SP MP2 6-31G*	-66.730	-25.842
	Het[D*H] CP	-34.373	-11.642
	Het[D*H] OPT MP2 6-31G*	-66.435	
4	Het[DH*] SP MP2 6-31G*	-133.694	-95.313
	Het[DH*] CP	-102.305	-81.092
	Het[DH*] OPT MP2 6-31G*	-133.232	

Figure 6.15b Subsets of pairs that mismatch in two positions. MP2 and HF

Fig. 6.22 Number key	2 Mismatches	MP2 TIE (kJ/mol)	HF TIE (kJ/mol)
6	Het[AD] SP MP2 6-31G*	-57.958	-22.73
	Het[AD] CP	-35.331	-11.796
	Het[AD] OPT MP2 6-31G*	-57.242	
7	Het[AG] SP MP2 6-31G*	-39.193	-4.379
	Het[AG] CP	-15.529	7.503
8	Het[BE] SP MP2 6-31G*	-90.614	-51.89
	Het[BE] CP	-62.718	-38.198
	Het[BE] OPT MP2 6-31G*	-90.156	
9	Het[AG*] SP MP2 6-31G*	-2.178	40.894
	Het[AG*] CP	22.575	52.766
10	Het[A*D*] SP MP2 6-31G*	75.895	120.041
	Het[A*D*] CP	103.047	131.782
11	Het[A*G*] SP MP2 6-31G*	28.346	77.844
	Het[A*G*] CP	55.368	89.439
12	Het[B*E*] SP MP2 6-31G*	37.937	86.288
	Het[B*E*] CP	68.563	99.906
13	Het[A*G] SP MP2 6-31G*	104.128	144.832
	Het[A*G] CP	136	159.096
14	Het[EH] SP MP2 6-31G*	-23.852	11.194
	Het[EH] CP	4.731	24.678
15	Het[DF*] SP MP2 6-31G*	-3.691	34.741
	Het[DF*] CP	28.337	49.654
16	Het[CH] SP MP2 6-31G*	24.155	60.133
	Het[CH] CP	52.547	73.455
17	Het[DF] SP MP2 6-31G*	1.428	45.968
	Het[DF] CP	29.007	58.936

Figure 6.15c Subsets of pairs that mismatch in three positions. MP2 and HF

Fig. 6.22 Number key	3 Mismatches	MP2 TIE(kJ/mol)	HF TIE (kJ/mol)
19	Het[AB] SP MP2 6-31G*	13.265	54.689
	Het[AB] CP	33.919	65.406
20	Het[AE] SP MP2 6-31G*	-24.249	10.491
	Het[AE] CP	-4.004	21.069
	Het[AE] OPT MP2 6-31G*	-24.35	
21	Het[A*B*] SP MP2 6-31G*	195.092	243.238
	Het[A*B*] CP	222.435	255.331
22	Het[A*E*] SP MP2 6-31G*	168.825	224.745
	Het[A*E*] CP	194.241	235.867
23	Het[C*D*] SP MP2 6-31G*	134.474	178.201
	Het[C*D*] CP	160.717	190.759
24	Het[CF*] SP MP2 6-31G*	60.196	93.246
	Het[CF*] CP	88.619	107.586
	Het[CF*] OPT MP2 6-31G*	58.489	
25	Het[B*D*] SP MP2 6-31G*	127.953	169.775
	Het[B*D*] CP	153.185	181.892
26	Het[E*G*] SP MP2 6-31G*	81.642	129.925
	Het[E*G*] CP	108.612	142.57
27	Het[CD] SP MP2 6-31G*	61.536	97.03
	Het[CD] CP	86.979	109.516
28	Het[C*F] SP MP2 6-31G*	91.74	131.486
	Het[C*F] CP	121.516	145.645
	Het[C*F] OPT MP2 6-31G*	91.491	
29	Het[BD] SP MP2 6-31G*	48.942	91.439
	Het[BD] CP	71.784	102.515
30	Het[EG] SP MP2 6-31G*	30.733	65.727
	Het[EG] CP	56.666	79.062

Table 6.15d Subsets of pairs that mismatch in four positions. MP2 and HF

Fig. 6.22 Number key	4 Mismatches	MP2 TIE(kJ/mol)	HF TIE (kJ/mol)
32	Het[AA] SP MP2 6-31G*	43.899	86.756
	Het[AA] CP	57.760	94.623
33	Het[BB] SP MP2 6-31G*	106.043	151.571
	Het[BB] CP	127.571	163.003
34	Het[CC] SP MP2 6-31G*	103.223	134.624
	Het[CC] CP	125.907	146.907
35	Het[DD] SP MP2 6-31G*	182.744	223.871
	Het[DD] CP	204.003	234.714
36	Het[FF] SP MP2 6-31G*	145.654	182.290
	Het[FF] CP	172.840	196.386

Table 6.15d(Continued) Subsets of pairs that mismatch in four positions, MP2 and HF

Fig. 6.22 Number key	4 Mismatches	MP2 TIE(kJ/mol)	HF TIE (kJ/mol)
37	Het[GG] SP MP2 6-31G*	183.943	217.341
	Het[GG] CP	210.766	231.550
38	Het[A*A*] SP MP2 6-31G*	329.362	385.747
	Het[A*A*] CP	351.719	395.652
39	Het[B*B*] SP MP2 6-31G*	241.837	285.200
	Het[B*B*] CP	267.556	297.768
40	Het[C*C*] SP MP2 6-31G*	211.276	258.720
	Het[C*C*] CP	235.900	270.638
41	Het[G*G*] SP MP2 6-31G*	140.103	190.479
	Het[G*G*] CP	165.908	203.717

A similar picture to that for complementary associations is evident here, MP2 in each case gives a lower (more binding) TIE. On average the energy difference between the HF and MP2 (over mismatching associations) is 42 kJ/mol. Removal of BSSE brings the average (mismatching associations) difference to 28 kJ/mol. This difference in TIE when the geometry is same for both methods can be attributed to correlation.

Examining the results graphically (Fig. 6.22) reveals that although the TIE values calculated with MP2 are lower then the corresponding HF the overall relative pattern of TIEs remains unchanged.

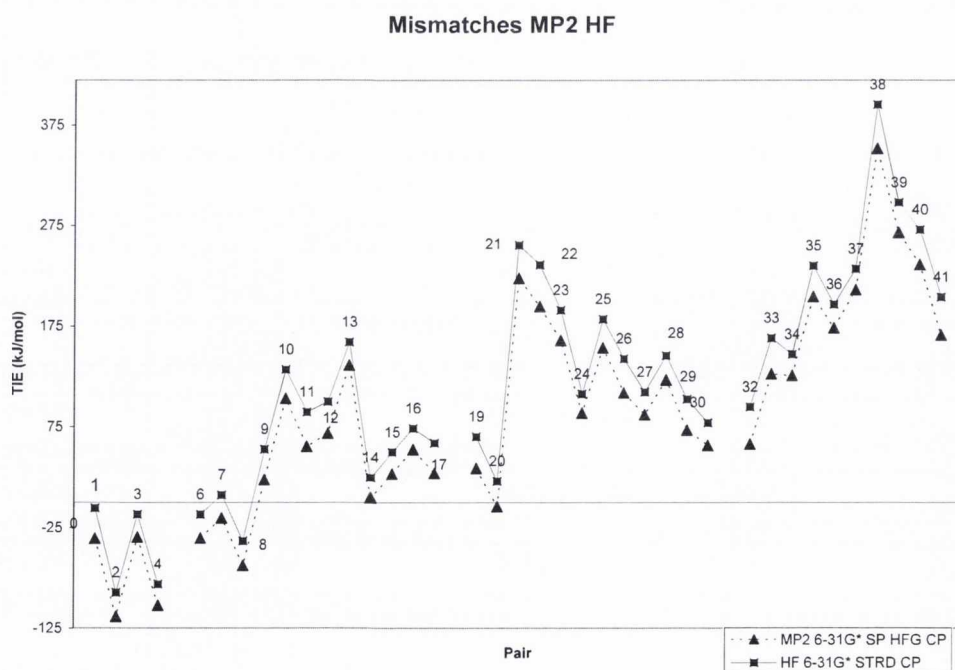


Figure 6.22 Subsets of pairs that mismatch in one, two and three. MP2 and HF

These results lead to the conclusion that performing calculations at the MP2 level to include electron correlation does not change the overall relative pattern of results seen, energy values are merely shifted down in value. As the shift in values seen is fairly uniform across the test data set, the results determined can be used to estimate the MP2 results for the complete Het data set.

6.9 Summary and Discussion – MP2

In the above section calculations were performed on 46 out of a possible 136 associations. The data collected from the MP2 tests can be used to predict the outcome of MP2 calculations on the entire Het set. An approximate correlation value was calculated by taking the difference between HF and MP2 results including CP and averaging this over all test pairs. The average difference was found to be 28 kJ/mol. This value was then used to re-plot HF results (Fig. 6.23).

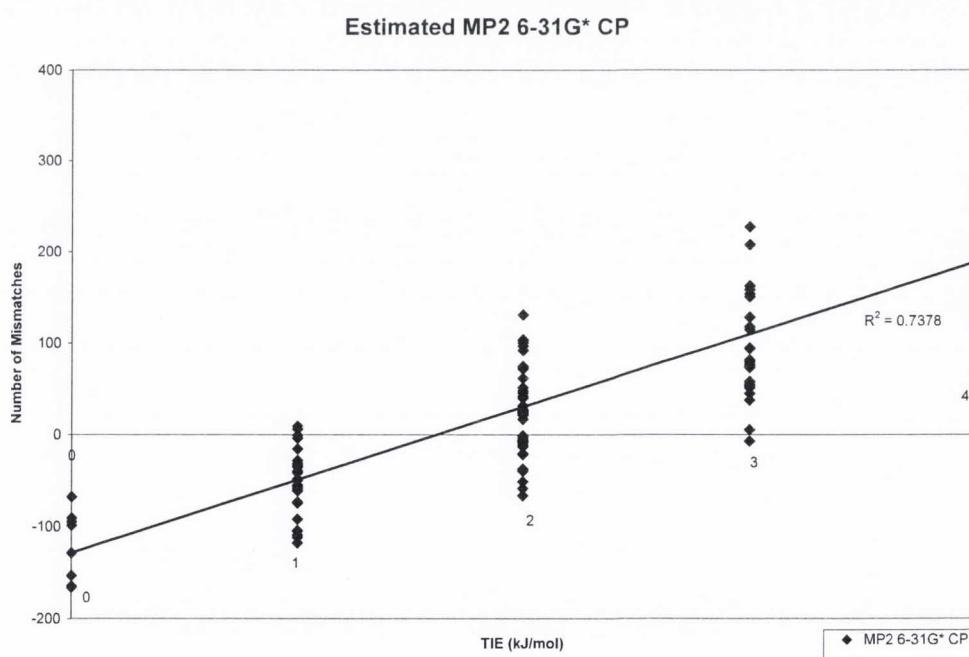


Figure 6.23 All Het pairs with estimated MP2 values

As the prediction implies that the results will only be shifted in value, the same trend previously seen for HF (section 5.3) is evident. Even with a shift of 28 kJ/mol, no change is seen in the number of surviving pairs. Three pairs Het[DD*], Het[FF*] and Het[GG*]

still remain that can energetically exist together as a group. The lowest most binding (limiting) pair in this group is 45 kJ/mol. This is large enough that even when the calculated average MP2 shift is subtracted the pairs are still repulsive.

The results for both semi-empirical methods and MP2 show changes in absolute values determined for individual total interaction energies between associations compared to the results set for HF calculations. Although changes in absolute values are evident between all the calculation methods considered the overall outcome remains the same irrespective of computational method used. This consistency in results seen suggests that the outcome is a consequence of the Het set of molecules and not simply a direct result of any particular computational method used. This exploration into the study of a complete set of molecules each with 4D/A patterns can be taken further by considering an alternate molecular representation of each pattern. A Zimmerman (Zim) based set of molecules will be used to further explore a 16 letter (4-bit) alphabet in the next section.

7 The Zimmerman Potential Alphabet

7.1 Introduction

It is important to determine if the results seen for the proposed Het alphabet are actually a result of the D/A patterns they convey. The particular choice of calculation method was considered in the previous chapter, however it is equally (if not more so) important to consider how the particular Het molecules chosen to represent the D/A patterns could have affected the results seen. In order to explore how this ideal Het set of molecules would compare to a less ideal alternate set of letters, a molecular alphabet was constructed based largely on work from the Zimmerman laboratory [1].

The Zimmerman [Zim] based alphabet like the Het one will consist of 16 molecules. In the construction of the Zim alphabet D/A patterns 0011-1100, 0101-1010, (Fig. 7.1(a, b)[1]) were taken from work by Zimmerman et al.. Pattern 0110-1001 (Fig. 7.1 (c)[1]) is used with a hydrogen as the R group. 0010-1101 as shown in (Fig. 7.1 (d)[2]) was used with the R group as NH₂ and the two methyl groups attached to the ring structure in the molecule with the pattern 0010 were substituted with hydrogen atoms.

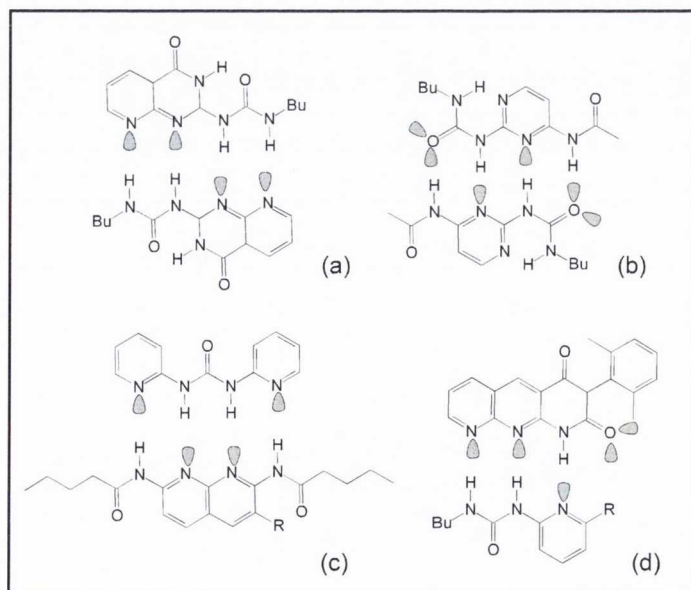


Figure 7.1 Pairs taken from the work of Zimmerman et al.

Not all of the pairs needed to form a full alphabet could be constructed based on the work of Zimmerman; examples of three pairs could not be found 0000-1111, 0001-1110, 0111-1000[3]. In order to create a complete set of 8 complementary pairs (that can be compared to the full set of Het pairs), the pairs for which examples could not be found need to be theoretically designed. This was done by building molecules similar in shape, size and flexibility to the literature members of the Zim set. The complete Zim potential alphabet set of letters is shown below (Fig. 7.2). In the complete set of Zim associations Zim[BB*] and Zim[HH*] use the same molecular configuration, similarly Zim[CC*] and Zim[EE*] do so also. This is possible due to the fact that in this initial exploration of the Zim set of letters an anchoring backbone structure is not in place. These simplifications along with others that can be made in the Zim set are set out in appendix A9. The reader should note that results will always be shown in full to give a complete and directly comparable data set of results. The reader should also note that since the completion of the work presented in this chapter a stable AAAA-DDDD quadruple hydrogen bonding array has been reported in the literature [4, 5].

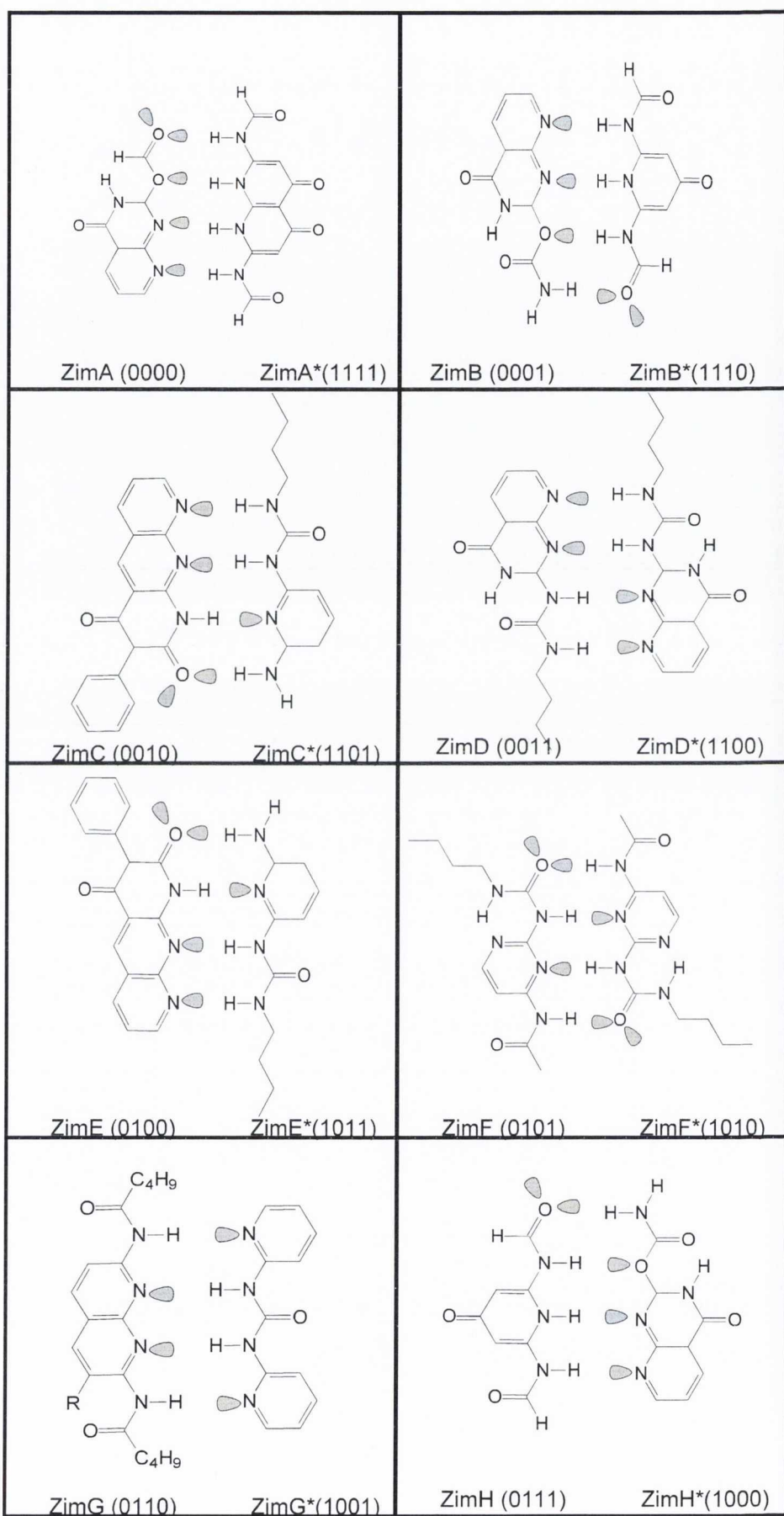


Figure 7.2 Zimmerman alphabet. 16 letters/molecules broken into 8 complementary pairs.

Comparing the Zim letters to the Het (Fig. 3.3) it is apparent that the molecules in the Zim alphabet are not all uniform in shape or structure. The number of ring structures in each molecule in the Zim potential alphabet set is not fixed; it varies from just one (as in ZimH) to as many as four (as in ZimC). Several of the molecules in the Zim set have long carbon chains in terminal positions. These long carbon chains and the variation in the number of ring structures gives the molecules different degrees of freedom and flexibility compared to the Het. This variation in structure could potentially lead to a greater degree of bending and twisting than what was seen for the Het potential letter set.

In this chapter the proposed Zim potential alphabet set will be explored and compared to the previously discussed Het set of letters. In exploring the Zim pairs a similar procedure to that adopted in the study of the Het set will be employed. First the free complementary associations will be considered and used to develop STRD molecular geometry restrictions, these STRD restrictions will then be used in the study of non-complementary associations. At each appropriate step the results determined will be compared to those for the corresponding Het association. If the results seen for the Het alphabet are due to the hydrogen bond D/A pattern and not intrinsically linked to the molecular structures used, the results seen for the Zim alphabet should show a similar final result.

7.2 Zimmerman Results

7.2.1 Free complementary associations

Calculations were carried out on the 8 complementary pairs using HF with a 6-31G(d) basis set. In the first instance no constraints other than keeping the C_s point group were in place. The full TIE results are shown below (Table 7.1).

Table 7.1 Complementary pairs TIE. Free HF 6-31G(d) basis set. C_s point group

	X1(au)	X2(au)	X1+X2	X1:X2 (au)	TIE(kJ/mol)
Zim[AA*] 0000-1111	-693.815	-900.582	-1594.397	-1594.429	-83.900
Zim[BB*] 0001-1110	-748.894	-657.087	-1405.981	-1406.020	-104.250
Zim[CC*] 0010-1101	-979.297	-680.706	-1660.003	-1660.023	-53.320
Zim[DD*] 0011-1100	-885.214	-885.214	-1770.428	-1770.464	-94.954
Zim[EE*] 0100-1011	-979.297	-680.706	-1660.003	-1660.023	-53.320
Zim[FF*] 0101-1010	-848.509	-848.509	-1697.019	-1697.055	-96.342
Zim[GG*] 0110-1001	-715.080	-1063.217	-1778.297	-1778.303	-16.445
Zim[HH*] 0111-1000	-657.087	-748.894	-1405.981	-1406.020	-104.250

As expected, all the pairs are binding. A large spread in TIEs is seen Zim[BB*], Zim[HH*] are the most attractive and Zim[GG*] the least. The results can be directly compared to those for the Het complementary pairs (Fig. 7.3).

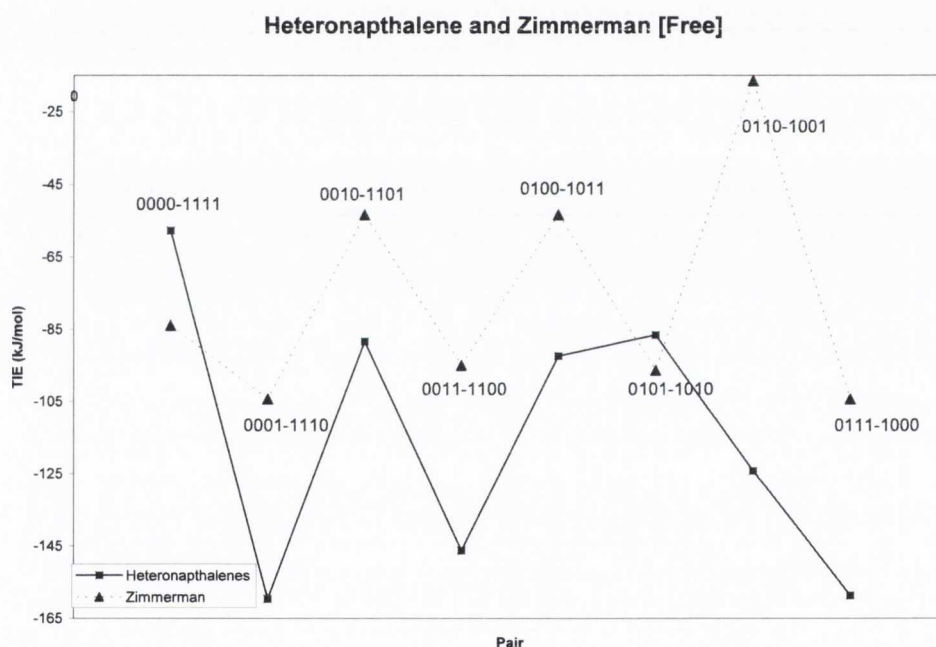


Figure 7.3 Complementary pairs TIE. Het and Zim

The pattern of relative energies seen in the Zim results deviates from that seen for the Het, in absolute values but also in shape. Three main points of disagreement are clear between the two molecular sets, AA* 0000-1111, FF* 0101-1010 and GG* 0110-1001 (If Zim or Het are not used in front of a letter name or D/A pattern, for example, GG* 0110-1001 this refers to the letters and their pattern and not a specific molecular representation of the letter with that pattern.).

In the Het alphabet pattern AA* 0000-1111 is the least binding pair. It is the only pair to contain an oxygen atom as part of the main ring structure, causing it to differ from the other pairs. This structural difference could also be responsible for causing the disagreement at this point between the two hypothetical alphabets. The pattern FF* 0101-1010 in the Het alphabet is one of the weakest binding pairs, whilst in the Zimmerman alphabet it is one of the strongest. This increase in TIE seen on comparing the two molecular sets could possibly in part be due to the less rigid structure that the FF* 0101-1010 pattern is represented by in the Zim set.

If this pair is forced into a more rigid structure (Fig. 7.4), a decrease in TIE of 28 kJ/mol is seen (calculated using AM1).

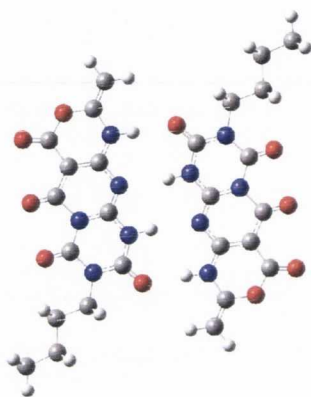


Figure 7.4 Alternative Zim[FF*] rigid structure.

The TIE of Het[FF*] and Zim[FF*] do not differ that greatly from each other (<10 kJ/mol). It may in fact not be Zim[FF*] that exhibits extra stability but rather the other Zim associations are too repulsive. An answer to this could not be determined without trying many examples of each Zim letter similar to each other in structure.

In the Zimmerman set pattern GG* 0110-1001 is shown to be the least binding in the set. This relatively low TIE is most probably due to the curved shape of GG* in the Zim. This curvature arises as a result of large subsistent groups causing the molecules to bend and sit much further apart in order to interact. If an alternative structure with a rigid ring structure (Fig. 7.5) were instead used (as the molecules are no longer free to bend), an increase of 25 kJ/mol is seen in the TIE (determined using AM1).

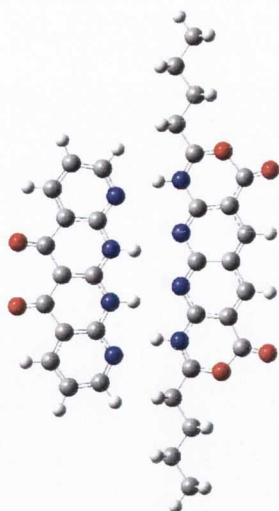


Figure 7.4 Alternative Zim[GG*] rigid structure.

It is worth noting that a large deviation in association constants has been reported by Zimmerman and Corbin [1] for complexes with the same D/A pattern but with different structures.

7.2.2 Standard geometry constraints

As was done previously for the full Het potential alphabet geometry constraints were devised to mimic the effect of a restricting pocket. Complementary associations were used to determine appropriate constraints for the full Zim potential set of letters. These STRD Zim constraints will then be used to model non-complementary associations. In establishing geometry constraints the same procedure (including computational method and basis set) as described in section 3.2.1 of analysing the optimized geometry of free complementary pairs was used. The results are presented in Table 7.2a and Table 7.2b.

Table 7.2a Geometry of complementary pairs optimised with HF 6-31G(d) basis set.

The four hydrogen bond distances in α β γ δ positions are measured for each pair

HF 6-31G* Free	α (Å)	β (Å)	γ (Å)	δ (Å)	Average $\beta\gamma$ (Å)
Zim[AA*] 0000-1111	3.12	3.65	3.62	3.21	3.64
Zim[BB*] 0001-1110	3.16	3.57	3.56	2.88	3.56
Zim[CC*] 0010-1101	3.39	3.69	3.51	2.97	3.60
Zim[DD*] 0011-1100	3.15	3.43	3.43	3.15	3.43
Zim[EE*] 0100-1011	2.97	3.51	3.69	3.39	3.60
Zim[FF*] 0101-1010	2.92	3.56	3.56	2.92	3.56
Zim[GG*] 0110-1001	4.00	4.63	4.63	4.00	4.63
Zim[HH*] 0111-1000	2.88	3.56	3.57	3.16	3.56
Average over all positions	3.20	3.70	3.70	3.21	3.70
Average $\beta\gamma$ excluding Zim[GG*]	3.08	3.57	3.56	3.10	3.56

Table 7.2b Geometry of complementary pairs optimised with HF 6-31G(d) basis set.

The four hydrogen bond angles α β γ δ positions are measured for each pair

HF 6-31G* Free	α (degrees)	β (degrees)	γ (degrees)	δ (degrees)	Average $\beta\gamma$ (degrees)
Het[AA*] 0000-1111	167.725	168.192	175.621	178.391	171.907
Het[BB*] 0001-1110	179.872	177.432	174.017	178.690	175.725
Het[CC*] 0010-0010	176.150	178.381	178.751	178.506	178.566
Het[DD*] 0011-1100	179.416	175.980	175.982	179.446	175.981
Het[EE*] 0100-1011	175.640	178.151	178.046	179.301	178.099
Zim[FF*] 0101-1010	178.854	175.806	175.806	178.856	175.806
Zim[GG*] 0110-1001	179.788	178.303	178.294	179.776	178.298
Zim[HH*] 0111-1000	178.689	173.662	177.123	179.913	175.393
Average over all positions	177.017	175.738	176.705	179.110	176.222

Due to its curved structure Zim[GG*] deviates from the other pairs, for this reason it is neglected in the standard geometry constraints determination process. The average bond length over β and γ excluding Zim[GG*] was determined to be 3.56Å. The angle for the middle two positions was frozen at 180°. Although a larger degree of wobble and thus deviation from 180° is evident in the Zim free pairs compared to the Het, 180° is still chosen as the angle constraint for the same reasons as discussed for the Het letters (section 3.2.1) and also to allow consistent results comparison between the two data sets.

The TIEs for the Zim complementary pairs were recalculated with the STRD restrictions/conditions in place (Table 7.3)(Fig. 7.6).

Table 7.3 TIEs for free and standard (STRD) conditions

Pair	TIE Free (kJ/mol)	TIE STRD (kJ/mol)
Zim[AA*] 0000-1111	-83.900	-81.795
Zim[BB*] 0001-1110	-104.250	-92.236
Zim[CC*] 0010-1101	-53.320	-47.461
Zim[DD*] 0011-1100	-94.954	-93.833
Zim[EE*] 0100-1011	-53.320	-47.461
Zim[FF*] 0101-1010	-96.342	-87.145
Zim[GG*] 0110-1001	-16.445	4.312
Zim[HH*] 0111-1000	-104.250	-92.236

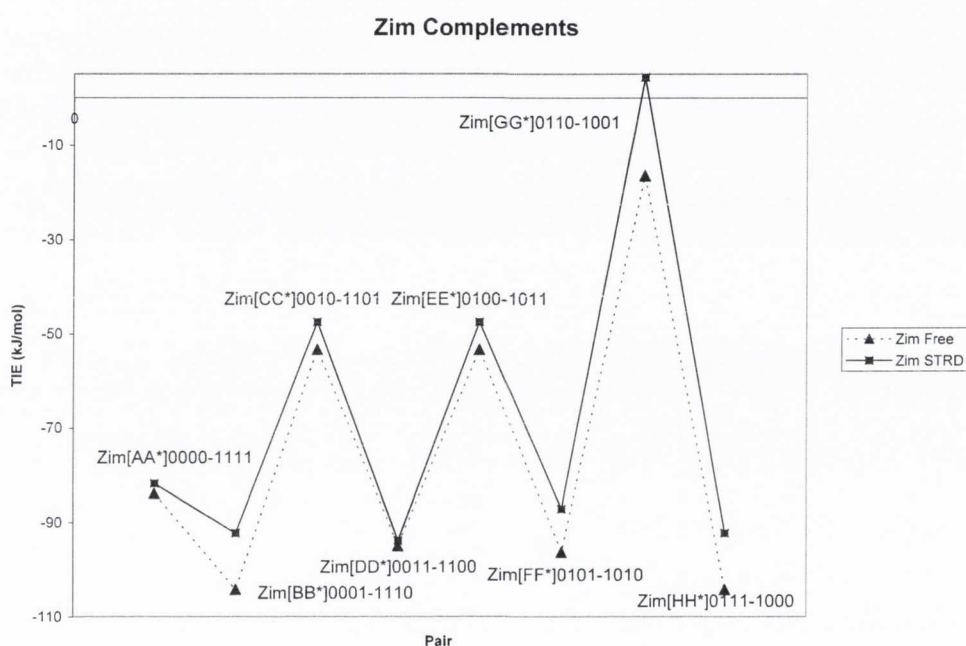


Figure 7.6 Plot of Zim TIEs for free and standard (STRD) conditions.

The individual interaction energies have changed (becoming more repulsive as the geometry is now restricted) but importantly the relative pattern of results remains the same. The largest change in energy is seen for Zim[GG*], this is to be expected as its geometry was furthest away from the other pairs and its bond lengths were excluded from the calculation of the STRD conditions. The TIE for Zim[GG*]; is above zero but as this energy is for a complementary association it will be considered to be binding. This is permissible in the case as the deviation in behaviour relates to molecule selection rather than D/A pattern. This result does indicate that the geometry of Zim[GG*] could potentially lead to other anomalous results and so its structure will be kept in mind during results analysis.

7.3 Non-complementary associations

In order to gain some insight into the potential viability of the proposed Zim set of letters all possible pairings need to be considered. The STRD conditions developed for the Zim potential set of letters will be used throughout and the C_s point group kept. In order to allow direct comparison with the Het set the same method (HF) and basis set (6-31G(d)) will be used.

7.3.1 Mismatches in one and three positions

The results for each mismatch group will first be presented for the Zim set and then compared to the Het set generally and finally in some detail. To aid comparison the results are ordered (where appropriate) from most attractive to least in each mismatch category based on the Het results.

The Zim results for the 32 associations that mismatch in one position (Table 7.4) reveal that even with one mismatching position an attractive interaction is still present (in almost all cases). No subset of pairs related by complementary letter can be found (each Lp-Lp mismatch, for example GH* 0110-1000, is related to a H-H mismatch by changing each letter to its complement G*H 1001-0111) in which both interactions are repulsive.

Table 7.4 TIEs for Zim mismatches in one position

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
Lp-Lp α XNOR 1000		H-H α XNOR 1000	
Zim[AH] 0000-0111	-55.662	Zim[A*H*] 1000-1111	16.532
Zim[DE] 0011-0100	-48.577	Zim[D*E*] 11-1011	9.463
Zim[CF] 0010-0101	-22.837	Zim[C*F*] 1101-1010	72.469
Zim[BG] 0001-0110	-9.371	Zim[B*G*] 1110-1001	0.134
Lp-Lp β XNOR 0100		H-H β XNOR 0100	
Zim[DH*] 0011-1000	-77.057	Zim[D*H] 1100-0111	-50.462
Zim[BF*] 0001-1010	-36.394	Zim[B*F] 1110-0101	-47.247
Zim[CG*] 0010-1001	-35.675	Zim[C*G] 1101-0110	31.897
Zim[AE*] 0000-1011	-12.445	Zim[A*E] 1111-0100	-39.214
Lp-Lp γ XNOR 0010		H-H γ XNOR 0010	
Zim[BD*] 0001-1100	-77.057	Zim[B*D] 1110-0011	-50.462
Zim[FH*] 0101-1000	-36.394	Zim[F*H] 1010-0111	-47.247
Zim[EG*] 0100-1001	-35.675	Zim[E*G] 1011-0110	31.897
Zim[AC*] 0000-1101	4.403	Zim[A*C] 1111-0010	-39.214
Lp-Lp δ XNOR 0001		H-H δ XNOR 0001	
Zim[CD*] 0010-1100	-48.577	Zim[C*D] 1101-0011	9.463
Zim[AB*] 0000-1110	-45.245	Zim[A*B] 1111-0001	16.532
Zim[EF*] 0100-1010	-22.837	Zim[E*F] 1011-0101	72.469
Zim[GH*] 0110-1000	-9.371	Zim[G*H] 1001-0111	0.134

Comparing the overall results of the Zim letters to the Het (Fig. 7.7), at some points similarities can be noted in the relative pattern of results although differences between the two data sets is evident at others. To explore these results further comparison plots were made for both letter sets broken down by the type of mismatch present and the position in which it occurs (Fig.7.8a-b)(Table 7.5).

One Mismatch (Het and Zim)

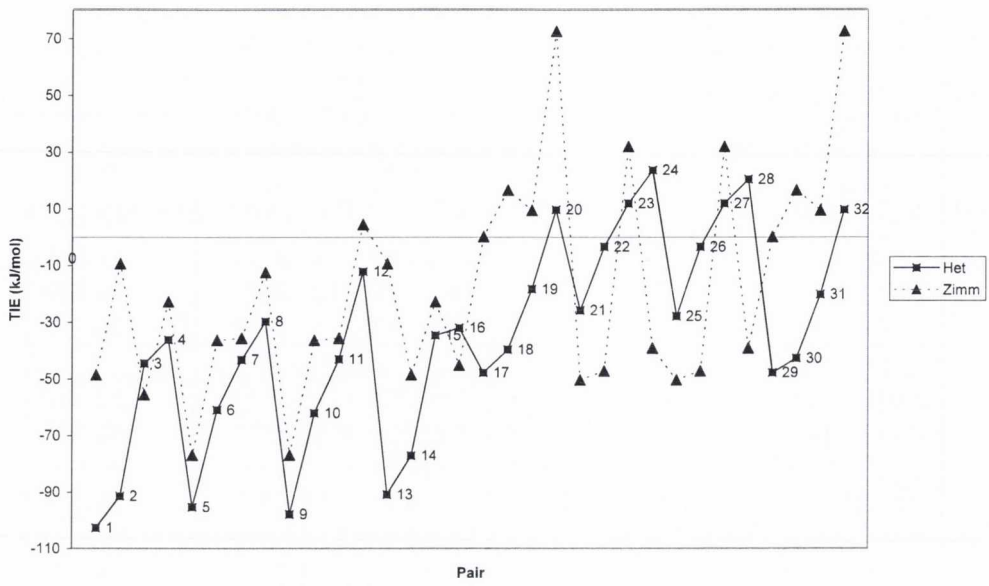


Figure 7.7 TIEs for mismatches in one position Het and Zim (number key Table. 7.5)

Mismatches in one position [Het and Zim]

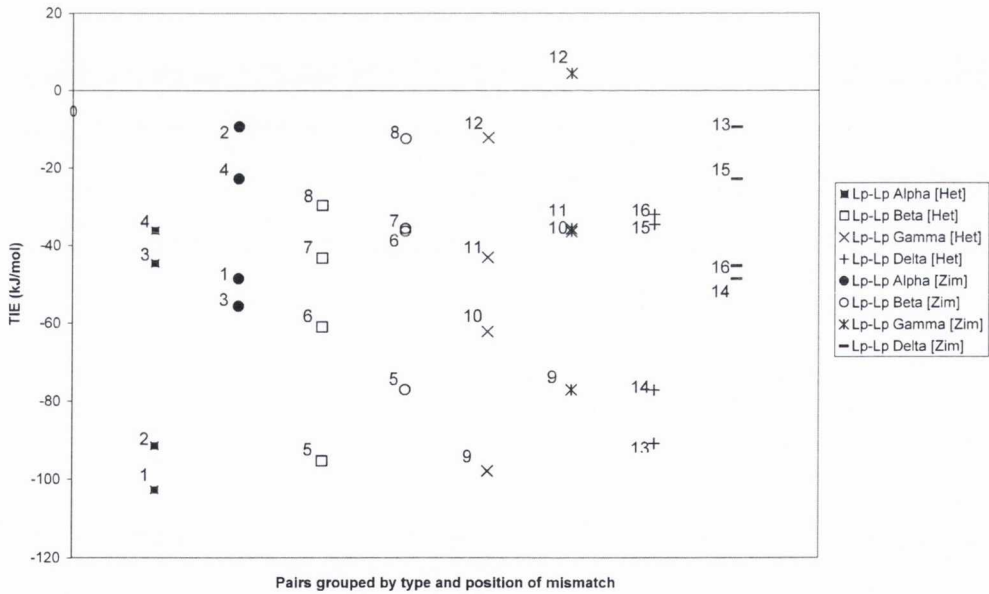


Figure 7.8a TIEs for Lp-Lp mismatches in one position Het and Zim (number key Table. 7.5)

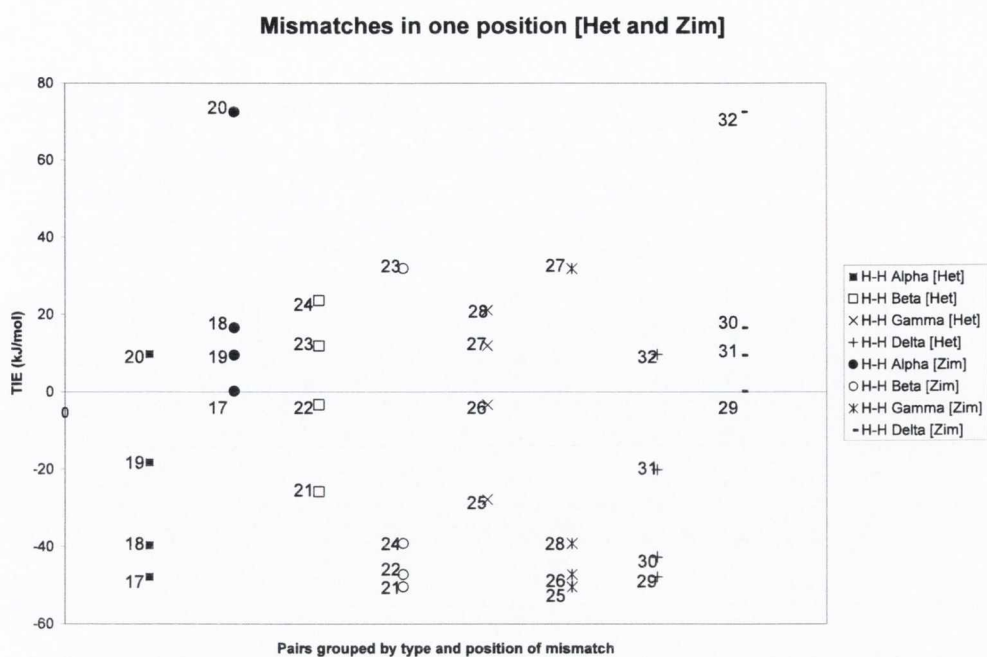


Figure 7.8b TIEs for H-H mismatches in one position Het and Zim (number key Table 7.5)

Table 7.5 Number key for mismatches in one position Het and Zim

	Lp-Lp α		H-H α
1	DE 0011-0100	17	B*G* 1110-1001
2	BG 0001-0110	18	A*H* 1111-1000
3	AH 0000-0111	19	D*E* 1100-1011
4	CF 0010-0101	20	C*F* 1101-1010
5	Lp-Lp β		H-H β
5	DH* 0011-1000	21	D*H 1100-0111
6	BF* 0001-1010	22	B*F 1110-0101
7	CG* 0010-1001	23	C*G 1101-0110
8	AE* 0000-1011	24	A*E 1111-0100
9	Lp-Lp γ		H-H γ
9	BD* 0001-1100	25	B*D 1110-0011
10	FH* 0101-1000	26	F*H 1010-0111
11	EG* 0100-1001	27	E*G 1011-0110
12	AC* 0000-1101	28	A*C 1111-0010
13	Lp-Lp δ		H-H δ
13	GH* 0110-1000	29	G*H 1001-0111
14	CD* 0010-1100	30	A*B 1111-0001
15	EF* 0100-1010	31	C*D 1101-0011
16	AB* 0000-1110	32	E*F 1011-0101

Examining the plot for Lp-Lp mismatches (Fig. 7.8a) reveals that just one out of the 16 of these associations $Zim[AC^*]$ is repulsive. As each Lp-Lp mismatch is related to a H-H by complement (Eg. $BD^* 0001-1100$ with one Lp-Lp mismatch and $B^*D 1110-0011$ with a H-H mismatch), Lp-Lp mismatches can be described as the limiting interactions for pairs that mismatch in one position. No matter how repulsive H-H interactions are they will be prevented from forming a viable alphabet due to their link with attractive Lp-Lp associations.

Analysis of the H-H mismatches (Fig. 7.8b) shows that many more of these associations are now repulsive in comparison to the Lp-Lp. Some deviations between the two sets of potential alphabets can be noted depending on whether the mismatch occurs in a terminal (alpha or delta) or middle (beta or gamma) D/A position. As STRD geometry constraints are only applied to the middle two positions it may seem intuitive that mismatches in these positions give the most repulsive energies. In the Zimmerman set a mismatch in a terminal position is on average more repulsive than a mismatch in a middle position, whilst in the Het set the middle two positions are on average more repulsive than the terminal two positions. The molecules in the Zim set are not all uniform or alike in structure, in contrast to the Het letters. These structural differences could be responsible for the change in behaviour noted. The difference in behaviour at the terminal and inner D/A positions of the two letters sets is most apparent at $Zim[C^*F^*] 1101-1010$ and $Zim[E^*F] 1011-0101$ (points 20 and 32 respectively), looking at these pairs in more detail reveals that $ZimC^*$ and $ZimE^*$ both contain a large terminal substituent group potentially resulting in steric interactions between the opposing hydrogen atoms when they are restricted to the C_s point group (Fig. 7.9).

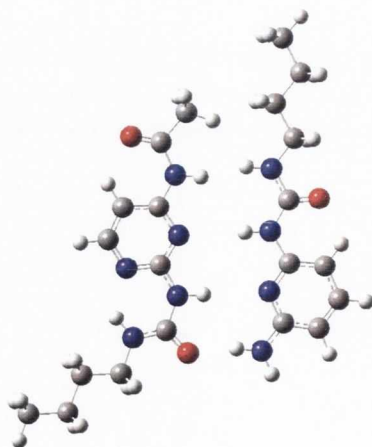


Figure 7.9 $Zim[F^*C^*] 1010-1101$

Zim[D*H] 1100-0111 and Zim[B*D] 1110-0011 (points 21 and 25) have the least repulsive energies of Zim pairs with one H-H mismatch. In both of these the mismatching position is not constrained to a rigid ring as it would be in the Het alphabet leading to a less repulsive interaction energy (Fig. 7.10).

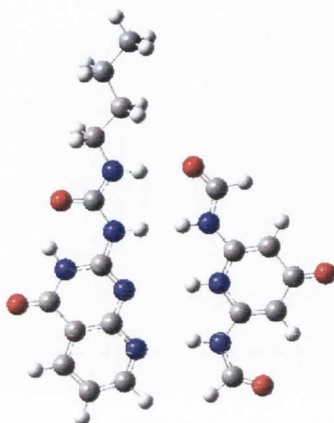


Figure 7.10 Zim[D*H*] 1100-0111

Even with the differences noted between the two molecular sets due to variation in geometry, the overall outcome for both of the molecular alphabets is the same: a mismatch in one position is not enough to ensure binding fidelity.

A link exists between pairs that mismatch in one and three positions (as previously discussed in 5.2.3). Any given pair that mismatches in one position (for example AB* 0000-1110) can be converted into a pair that mismatches in three positions by changing either of the molecules in a pair that mismatches in one position to its complement. The same applies to converting pairs that mismatch in three positions into pairs that mismatch in just one. Due to the two mismatch data sets being related in this way mismatches in three positions will be considered next.

Looking at the TIE results (Table 7.6) three mismatches is enough to make all but one of the associations repulsive. The overall plot (Fig. 7.11) for both molecular sets shows that the two sets follow different results patterns with the Het results showing less variation between consecutive data points. The results can be analysed further further by breaking the associations down in groups based on mismatch type (Fig. 7.12).

Table 7.6 TIEs for Zim mismatches in one position

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
XNOR 0111		XNOR 1101	
Zim[AH*] 0000-1000	9.022	Zim[AC] 0000-0010	0.063
Zim[D*E] 1100-0100	46.126	Zim[EG] 0100-0110	56.273
Zim[CF*] 0010-1010	48.102	Zim[F*H*]1010-1000	80.042
Zim[BG*] 0001-1001	71.578	Zim[BD] 0001-0011	109.646
Zim[C*F] 1101-0101	77.257	Zim[E*G*] 1011-1001	184.831
Zim[B*G] 1110-0110	71.881	Zim[FH] 0101-0111	114.747
Zim[DE*] 0011-1011	163.607	Zim[B*D*]1110-1100	107.696
Zim[A*H] 1111-0111	145.750	Zim[A*C*] 1111-1101	154.645
XNOR 1011		XNOR 1110	
Zim[AE] 0000-0100	21.227	Zim[AB] 0000-0001	-12.326
Zim[BF] 0001-0101	80.042	Zim[EF] 0100-0101	48.102
Zim[D*H*] 1100-1000	109.646	Zim[CD] 0010-0011	46.126
Zim[CG] 0010-0110	56.273	Zim[G*H*]1001-1000	71.578
Zim[C*G*] 1101-1001	184.831	Zim[E*F*] 1011-1010	77.257
Zim[B*F*] 1110-1010	114.747	Zim[GH] 0110-0111	71.881
Zim[DH] 0011-0111	107.696	Zim[C*D*] 1101-1100	163.607
Zim[A*E*] 1111-1011	154.645	Zim[A*B*] 1111-1110	145.750

Three Mismatches (Het and Zim)

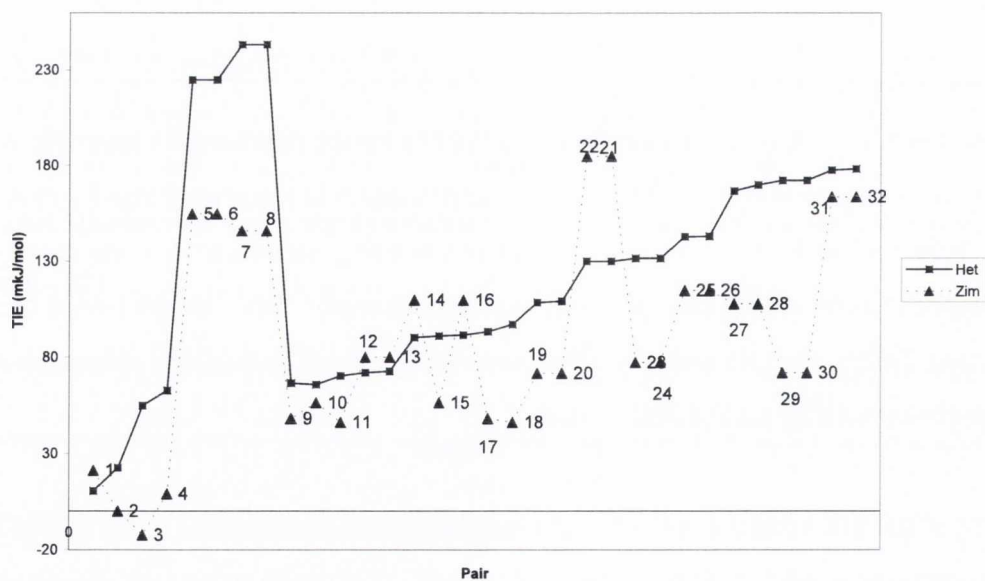


Figure 7.11 TIEs for mismatches in three positions Het and Zim (number key Table. 7.7).

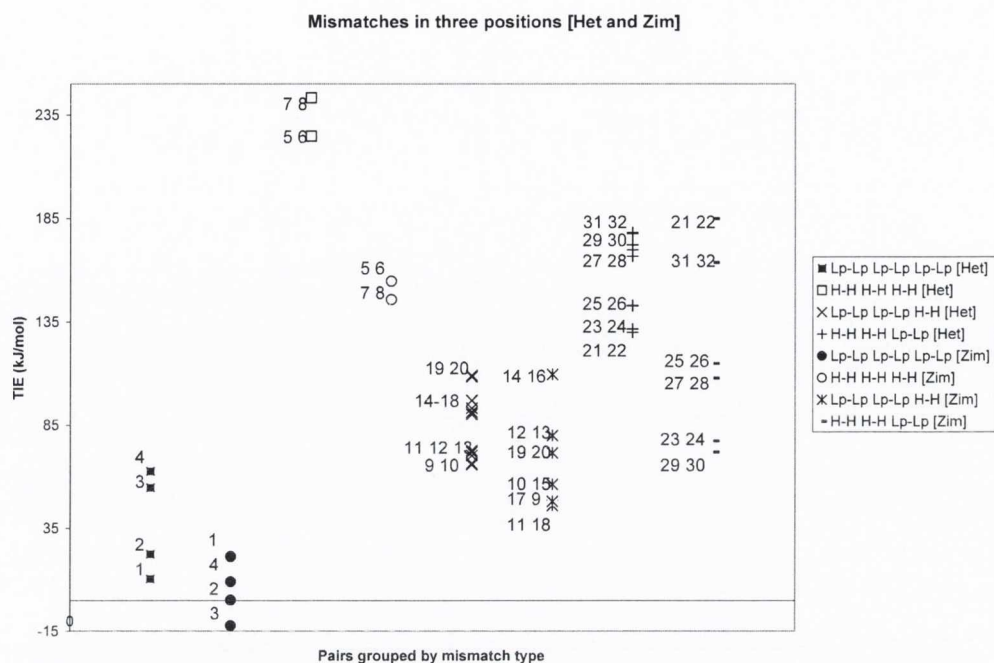


Figure 7.12 TIEs for H-H mismatches in three positions Het and Zim (number key Table.7.7)

Table 7.7 Number key for mismatches in three positions Het and Zim

1	Lp-Lp Lp-Lp Lp-Lp AE 0000-0100		
2	AC 0000-0010		
3	AB 0000-0001		
4	AH* 0000-1000		
5	H-H H-H H-H A*C* 1111-1101		
6	A*E* 1111-1011		
7	A*B* 1111-1110		
8	A*H 1111-0111		
9	Lp-Lp Lp-Lp H-H EF 0100-0101	21	H-H H-H Lp-Lp C*G* 1101-1001
10	EG 0100-0110	22	E*G* 1011-1001
11	D*E 1100-0100	23	E*F* 1011-1010
12	BF 0001-0101	24	C*F 1101-0101
13	F*H* 1010-1000	25	B*F* 1110-1010
14	D*H* 1100-1000	26	FH 0101-0111
15	CG 0010-0110	27	DH 0011-0111
16	BD 0001-0011	28	B*D* 1110-1100
17	CF* 0010-1010	29	B*G 1110-0110
18	CD 0010-0011	30	GH 0110-0111
19	G*H* 1001-1000	31	DE* 0011-1011
20	BG* 0001-1001	32	C*D* 1101-1100

The results when broken down by mismatch type (Fig. 7.12) show that when three mismatches are present that are all equal in type (Lp-Lp Lp-Lp Lp-Lp or H-H H-H H-H) the Zim pairs are less repulsive compared to the Het but similar in the relative patterns

seen. When a mixture of mismatch types is present the Zim pairs show a larger spread of energy values compared to the Het. This is most likely due to the increased variation in shape and flexibility in the Zim potential letter set. C*G* 1101-1001 and E*G* 1011-1001 (points 21 and 22) for example differ in relative repulsiveness on comparison of the Het and Zim sets. Their high repulsiveness in the Zim set is plausibly due to the curved structure of ZimG*. Zim[B*G] 1110-0110 and Zim[GH] 0110-0111 (points 29 and 30) show increased relative attraction compared to the corresponding Het associations, this is most probably a consequence of the increased flexibility of ZimB*, ZimG* and ZimH compared to the same Het letters.

As mismatches in one position (which cannot occur independently of mismatches in three positions) have already been ruled out for the Zimmerman alphabet, mismatches in three positions must also be ruled out from any potentially viable alphabet set. This overall result agrees with that of the Het potential letter set.

7.3.2 Mismatches in two positions

The 48 pairs with two mismatches can be broken into categories based on parity (even or odd) and complementary letter association. Doing this leads to the formation of 6 groups (in each parity category), each containing four pairs. The TIE results (Table. 7.8a, 7.8b) reveal a mixture of binding and repulsive energies.

Table 7.8a TIE for Zim mismatches in two positions with Even parity

Pair EVEN	TIE(kJ/mol)	Pair EVEN	TIE(kJ/mol)
XNOR 1100		XNOR 0110	
Zim[AD] 0000-0011	-20.183	Zim[DF] 0011-0101	95.427
Zim[A*D*] 1111-1100	56.377	Zim[D*F*] 1100-1010	95.427
Zim[AD*] 0000-1100	5.09	Zim[DF*] 0011-1010	-2.243
Zim[A*D] 1111-0011	56.377	Zim[D*F] 1100-0101	-2.243
XNOR 0101		XNOR 0101	
Zim[AF] 0000-0101	12.689	Zim[DG] 0011-0110	47.453
Zim[A*F*] 1111-1010	55.465	Zim[D*G*] 1100-1001	93.355
Zim[AF*] 0000-1010	-6.812	Zim[DG*] 0011-1001	93.355
Zim[A*F] 1111-0101	55.465	Zim[D*G] 1100-0110	47.453
XNOR 0110		XNOR 0011	
Zim[AG] 0000-0110	15.975	Zim[FG] 0101-0110	71.948
Zim[A*G*] 1111-1001	105.584	Zim[F*G*] 1010-1001	108.352
Zim[AG*] 0000-1001	-3.691	Zim[FG*] 0101-1001	108.352
Zim[A*G] 1111-0110	51.282	Zim[FG] 1010-0110	71.948

Table 7.8b TIE for Zim mismatches in two positions with Odd parity

Pair ODD	TIE(kJ/mol)	Pair ODD	TIE(kJ/mol)
XNOR 1100		XNOR 0110	
Zim[BC] 0001-0010	5.223	Zim[CE] 0010-0100	-18.763
Zim[B*C*] 1110-1101	49.942	Zim[C*E*] 1101-1011	100.59
Zim[BC*] 0001-1101	31.168	Zim[CE*] 0010-1011	33.669
Zim[B*C] 1110-0010	-12.352	Zim[C*E] 1101-0100	33.669
XNOR 0101		XNOR 0101	
Zim[BE] 0001-0100	-13.364	Zim[CH] 0010-0111	1.407
Zim[B*E*] 1110-1011	50.059	Zim[C*H*] 1101-1000	102.208
Zim[BE*] 0001-1011	102.208	Zim[CH*] 0010-1000	-13.364
Zim[B*E] 1110-0100	1.407	Zim[C*H] 1101-0111	50.059
XNOR 0110		XNOR 0011	
Zim[BH] 0001-0111	59.265	Zim[EH] 0100-0111	-12.352
Zim[B*H*] 1110-1000	59.265	Zim[E*H*] 1011-1000	49.942
Zim[BH*] 0001-1000	-61.48	Zim[EH*] 0100-1000	5.223
Zim[B*H] 1110-0111	31.209	Zim[E*H] 1011-0111	49.942

Looking at the overall pattern of results for both letter sets differences can be seen in the absolute value of energies and in their relative pattern (Fig. 7.13a)(Fig. 7.13b). It is noted that the total range of energy values for the two sets is similar (Fig. 7.14).

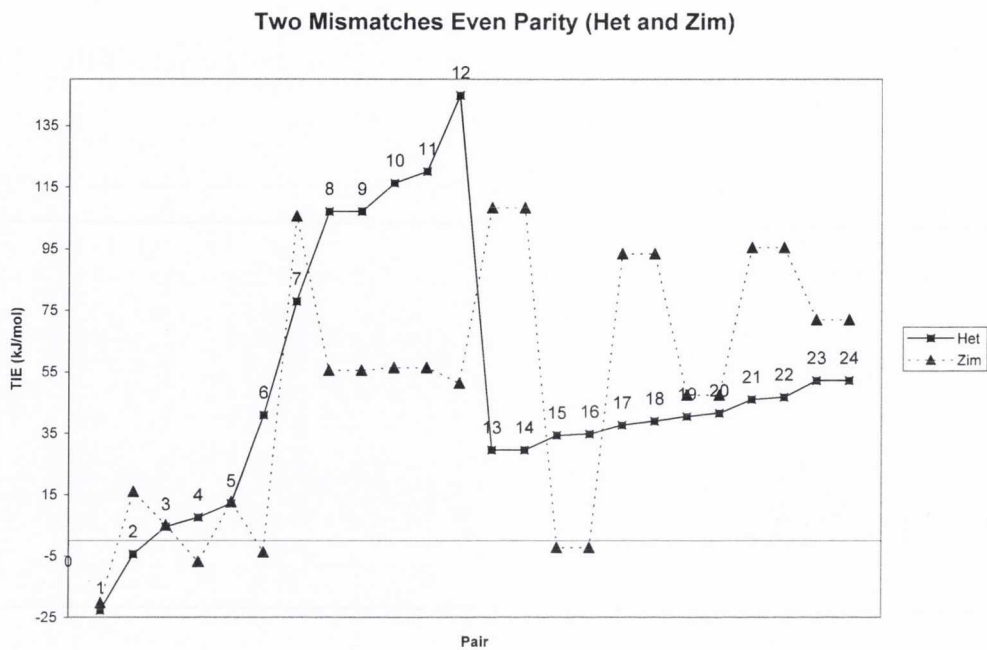


Figure 7.13a TIEs for mismatches in two positions Het and Zim (number key Table 7.9) Even Parity

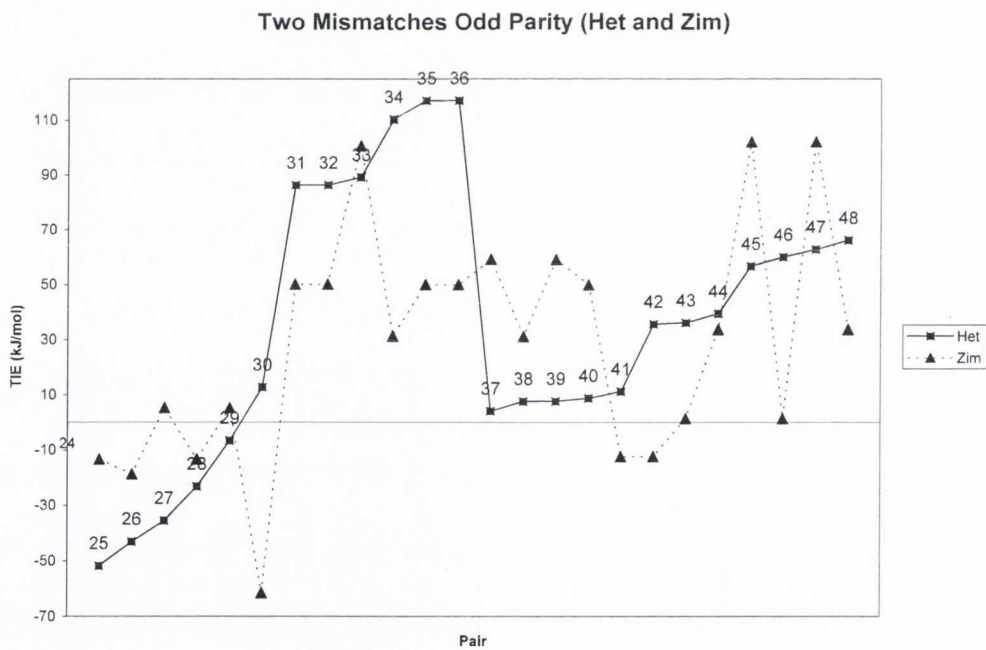


Figure 7.13b TIEs for mismatches in two positions Het and Zim (number key Table 7.9) Odd Parity

Mismatches in two positions Het and Zim

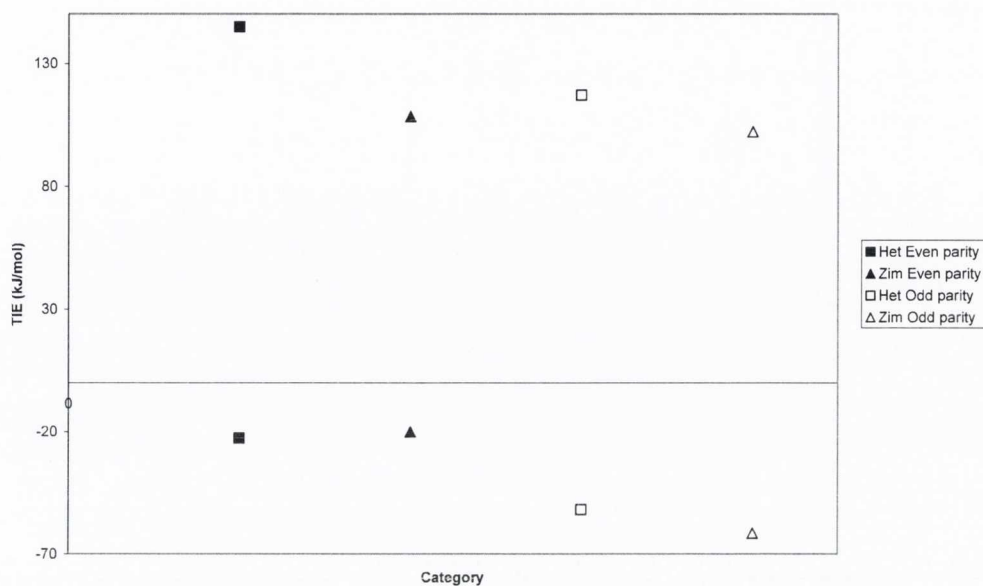


Figure 7.14 Range in TIEs for mismatches in two positions Het and Zim broken into parity sets.

Examining the pairs broken down by mismatch type (as well as parity) shows that overall the two data sets have a different relative pattern of results (Fig.7.15a)(Fig. 7.15b).

Mismatches in two positions [Het and Zim] Even

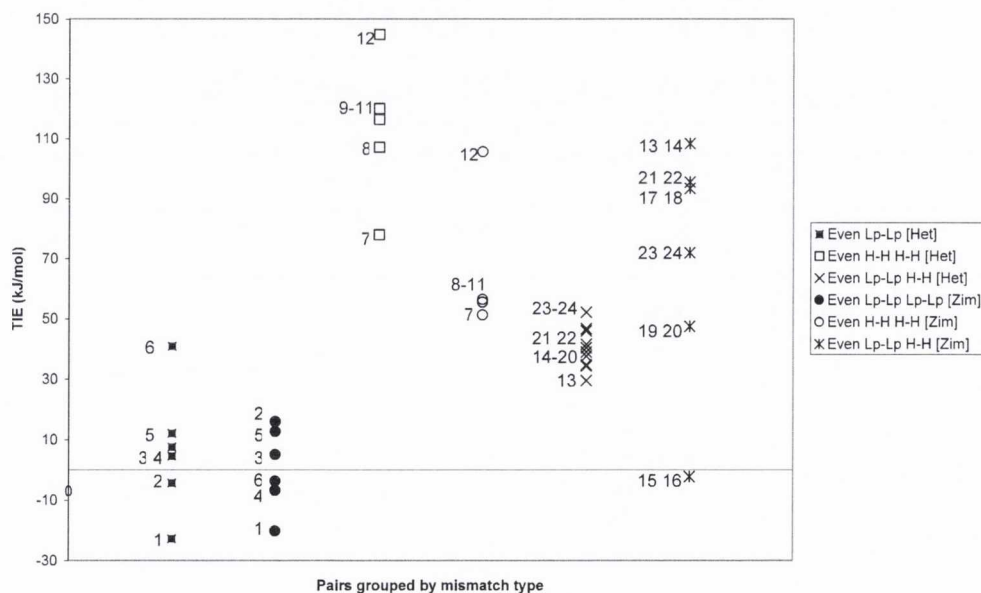


Figure 7.15a TIEs for mismatches in two positions Het and Zim (number key Table 7.9) Even Parity broken into sets based on mismatch type.

Mismatches in two positions [Het and Zim] Odd

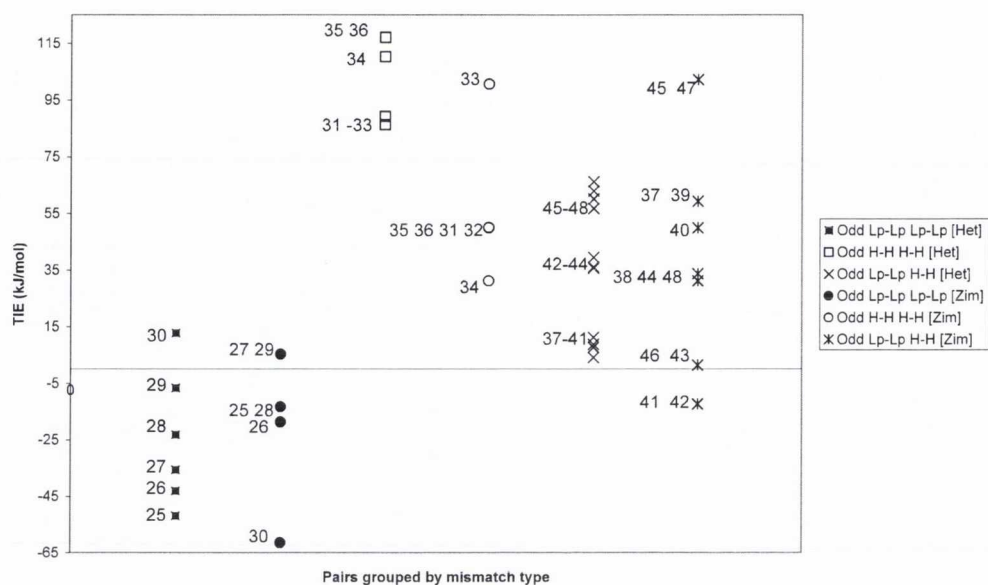


Figure 7.15b TIEs for mismatches in two positions Het and Zim (number key Table 7.9) Odd Parity broken into sets based on mismatch type

Table 7.9 Number key for mismatches in two positions Het and Zim

Even Parity		Odd Parity	
1	AD 0000-0011	25	BE 0001-0100
2	AG 0000-0110	26	CE 0010-0100
3	AD* 0000-1100	27	EH* 0100-1000
4	AF* 0000-1010	28	CH* 0010-1000
5	AF 0000-0101	29	BC 0001-0010
6	AG* 0000-1001	30	BH* 0001-1000
7	A*G* 1111-1001	31	C*H 1101-0111
8	A*F 1111-0101	32	B*E* 1110-1011
9	A*F* 1111-1010	33	C*E* 1101-1011
10	A*D 1111-0011	34	B*H 1110-0111
11	A*D* 1111-1100	35	B*C* 1110-1101
12	A*G 1111-0110	36	E*H 1011-0111
13	F*G* 1010-1001	37	BH 0001-0111
14	FG* 0101-1001	38	BC* 0001-1101
15	D*F 1100-0101	39	B*H* 1110-1000
16	DF* 0011-1010	40	E*H* 1011-1000
17	D*G* 1100-1001	41	EH 0100-0111
18	DG* 0011-1001	42	B*C 1110-0010
19	DG 0011-0110	43	B*E 1110-0100
20	D*G 1100-0110	44	C*E 1101-0100
21	DF 0011-0101	45	C*H* 1101-1000
22	D*F* 1100-1010	46	CH 0010-0111
23	FG 0101-0110	47	BE* 0001-1011
24	F*G 1010-0110	48	CE* 0010-1011

In both data sets associations with two H-H mismatches are clearly seen as more repulsive than Lp-Lp. Where mismatches are both of the same type overall Zim pairs show a slightly less repulsive range of energies compared to the corresponding Het associations. Where mismatches are mixed in type a larger range of values is evident for the Zim letters.

In order for any subset of pairs to have the possibility of coexisting all interaction other than those that are complementary must be repulsive. The results of the even parity (Table 7.8a) pairs reveal two sets (each containing two pairs) of letters, {ZimD, ZimD*, ZimG, ZimG*} {ZimF, ZimF*, ZimG, ZimG*} in which all non-complementary interactions are strongly repulsive (highlighted in black). In order for these two sets of letters to form a larger group containing all four pairs (as was possible in the case of the Het alphabet) four further interactions (highlighted in black) Zim[DF], Zim[D*F*], Zim[DF*], Zim[D*F] would need to be included. Two of these interactions Zim[DF], Zim[D*F*] are strongly repulsive whilst two Zim[DF*], Zim[D*F] are weakly attractive. Although an attraction of -2 kJ/mol is weak, it is still far more attractive than the energies shown by the other non-complementary associations in the group being considered. In the odd parity group of mismatches no set of pairs can be found in which all of the associations are repulsive (Table 7.8b). The results indicate that in the Zim set of letters two potentially viable sub-sets of pairs exists each containing four letters.

Comparing the results to those for the corresponding Het associations a similar overall picture is evident. The only potentially viable sub-set of Het letters is {HetD, HetD*, HetF, HetF, HetG, HetG*}, all of these letters except for F 0101 and F* 1010 are also members of the Zim potential set.

The important factor again noted here is that although changing the molecules used to represent individual D/A patterns does change the absolute values of the TIEs, the overall result remains unchanged.

7.3.3 Mismatches in four positions

To complete the study of the proposed Zim alphabet mismatches in all four possible positions were explored (Table 7.10).

Table 7.10 TIE for Zim pairs with four mismatches

Pair	TIE(kJ/mol)	Pair	TIE(kJ/mol)
Zim[AA] 0000-0000	30.307	Zim[A*A*] 1111-1111	265.098
Zim[BB] 0001-0001	124.690	Zim[B*B*] 1110-1110	181.277
Zim[CC] 0010-0010	63.574	Zim[C*C*] 1101-1101	238.764
Zim[DD] 0011-0011	179.082	Zim[D*D*] 0011-0011	179.082
Zim[EE] 0010-0010	63.574	Zim[E*E*] 1101-1101	238.764
Zim[FF] 0101-0101	232.432	Zim[F*F*] 0101-0101	232.432
Zim[GG] 0110-0110	155.229	Zim[G*G*] 1001-1001	347.985
Zim[HH] 0001-0001	124.690	Zim[H*H*] 1000-1000	124.690

The TIE for each of these pairs is as expected repulsive. The least repulsive TIE is shown by Zim[AA], due to it having only Lp-Lp repulsions. Zim[G*G*] has the most repulsive interaction energy. The large repulsive interaction energy found for this pair is most likely not a result of the H-H mismatches (in the terminal positions) but rather as a consequence of steric interactions between the large chains at either end of the molecule. When the complementary associations were examined Zim[GG*] was the furthest apart and deviated most from the average matched pair geometry. Forcing Zim[G*G*] to remain constricted to these standard conditions will cause the pair to be artificially too repulsive. This would be lessened if the standard conditions were relaxed and the pair given more freedom (for example anchored in space but free to twist).

Comparing the results of the two molecular potential alphabet sets (Fig. 7.16) overall similarities can be noted in the general trend. The most notable difference between the two set is at G*G* 1001-1001 for the molecular reason discussed above.

Mismatches in four positions [Het and Zim]

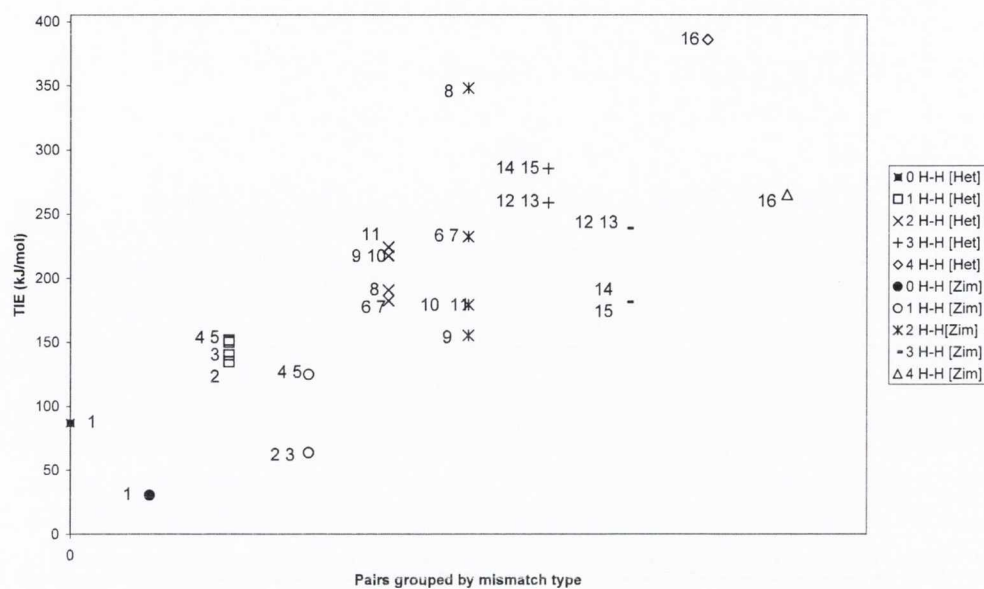


Figure 4.16 TIEs for Zim and Het pairs with four mismatches (number key Table 4.11)

Table 4.11 Number key for mismatches in four positions Het and Zim

1	0 H-H AA 0000-0000
2	1 H-H CC 0010-0010
3	EE 0100-0100
4	H*H* 1000-1000
5	BB 0001-0001
6	2 H-H F*F* 1010-1010
7	FF 0101-0101
8	G*G* 1001-1001
9	GG 0110-0110
10	DD 0011-0011
11	D*D* 1100-1100
12	3 H-H C*C* 1101-1101
13	E*E* 1011-1011
14	HH 0111-0111
15	B*B* 1110-1110
16	4 H-H A*A* 1111-1111

7.4 Summary and Discussion-Zimmerman results

The outcome of the study on the complete set of associations that can be formed between letters of the Zim potential alphabet set leads to a very similar overall conclusion as that derived for the Het letter set.

Deviation between the Het and Zim sets in both individual values and the relative ordering of values is observed. This deviation in absolute values is to be expected as the Het alphabet has been designed to act as an ideal model and is more rigid and uniform in structure compared to the Zim molecules. As the Zim set of molecules is less rigid in structure and more structural deviation exists between molecules in the set, the TIEs calculated for this potentially letter set could potentially be influenced by the particular molecules used. Effects such as steric interactions that will arise on combining some molecules (especially when as STRD geometry restrictions are in place) could potentially result in deviation from the expected pattern behaviour. Individual interaction energies are not central to comparing the two potential alphabets which due to their structural differences will never be identical in energy. In order to determine if the results are in fact linked to the D/A patterns (not just to the individual molecules chosen but to the patterns they in fact represent) the overall outcome of the comparison of the two sets is the pivotal factor.

For Zim mismatches in one position fidelity cannot be assured as most of these interactions are still binding. Mismatches in one position are related to those that mismatch in three positions. One type can be converted to the other by swapping one molecule of the pair to its complement (which must always be part of a possibly viable alphabet set to ensure the desired outcome of complementary pair formation). This link means that in order to prevent non-complementary binding associations both of these mismatch sets must be removed. Several of the associations that mismatch in two positions still exhibit a net attractive interaction energy. If the pairs are grouped into sets based on complementary association only two sets can be found in which all of the pairings are repulsive {ZimD, ZimD*, ZimG, ZimG*} {ZimF, ZimF*, ZimG, ZimG*}. These results are reflected in the plot below, number of mismatches versus the TIE (Fig. 7.17)

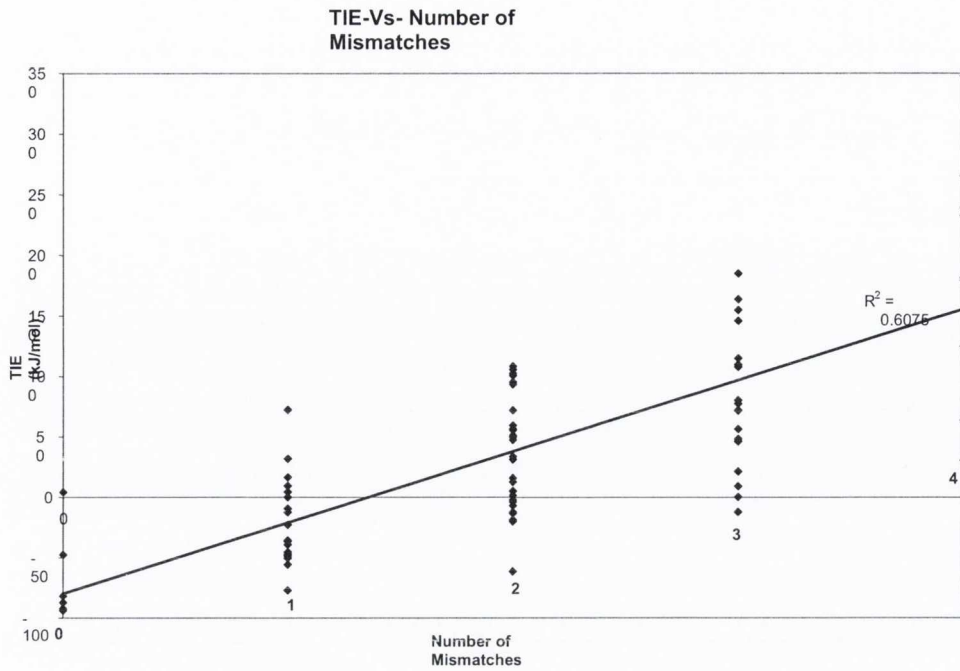


Figure 7.17 TIE versus number of mismatches all Zim associations

This plot is similar to that seen for the Het alphabet. The larger R^2 value of 0.7466 seen in the Het plot (Fig.5.13) indicates a larger overall spread of results is present in the Zim alphabet.

The range of TIE values seen for each mismatch is not unique to that mismatch, a large overlap between mismatch sets is evident. Mismatches can be grouped into sets based on complements. The lowest most binding pair in each set is known as the limiting pair. Looking at the plot of these limiting pairs only, a range of values can still be seen at each distinct mismatch number. Although the spread of values is less when examining limiting pairs only a significant overlap in TIE values is still present between the varying set of data for a given mismatch number (Fig. 7.18).

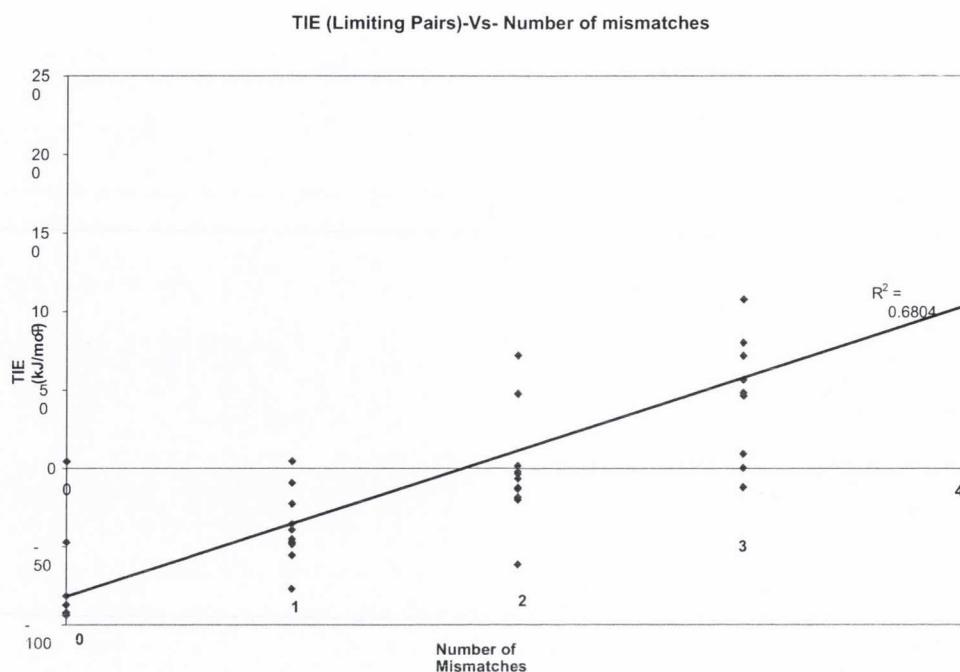


Figure 7.18 TIE versus number of mismatches for limiting Zim associations

In this plot of the Zim limiting associations the R^2 value is again lower than the corresponding Het value of 0.8769 meaning a larger spread of values is evident in the Zim associations.

In the exploration of the possible pairings in the Zim alphabet BSSE has not been calculated for each individual pair. Its inclusion would shift all pairs to more repulsive and thus less binding energies but would not change the overall outcome of the study. In the Zim potential alphabet all molecules are considered together as one large alphabet group. This leads to the Zim letters being inherently disadvantaged compared to an alphabet such as the nucleotide, in which both members of a complementary pair never exist in the same size group. If the Zim letter set was partitioned in some way, for example size, it would allow a greater minimum distance (δ) to be achieved between certain combination of letters and thus perhaps increasing the size of a potentially viable set of letters.

Having explored two potential alphabet letter sets, one ideal and one based where possible on literature and seeing the same overall result indicates that the results are in fact a consequence of the interactions between the D/A patterns and are not simply a consequence of the molecules used. With this important outcome established it is now

appropriate to take further considerations such as the potential effects of molecular flexibility into account.

1. Zimmerman, S.C. and F.S. Corbin, *Heteroaromatic modules for self-assembly using multiple hydrogen bonds*. Molecular Self-Assembly, 2000. **96**: p. 63-94.
2. Quinn, J.R. and S.C. Zimmerman, *With regard to the hydrogen bonding in complexes of pyridylureas, less is more. A role for shape complementarity and CH...O interactions?* Organic Letters, 2004. **6**(10): p. 1649-1652.
3. Wilson, A.J., *Non-covalent polymer assembly using arrays of hydrogen-bonds*. Soft Matter, 2007. **3**(4): p. 409-425.
4. Blight, B.A., et al., *An AAAA-DDDD quadruple hydrogen-bond array*. Nature Chemistry, 2011. **3**(3): p. 244-248.
5. Wilson, A.J., *Hydrogen bonding Attractive arrays*. Nature Chemistry, 2011. **3**(3): p. 193-194.

8 The effects of geometric freedom and flexibility

8.1 Introduction

In previous chapters exploring the Het and Zim sets of letters, total interaction energy values were used to assess potential viability. In these explorations all possible pairings for both the Het and Zim sets were modelled using geometrical restrictions (STRD conditions as discussed in 3.2.1) to approximate a hypothetical constrained environment. In accordance with the conditions necessary for viability (see section 1.4) any non-complementary association without a significantly repulsive TIE cannot be allowed to remain in any potential alphabet. Consider for example all possible interactions between the monomers HetB 0001, HetB* 1110, HetD 0011 and HetD* 1100 as shown in Table 8.1: these results indicate that any potential alphabet cannot contain both Het[BB*] and Het[DD*] as two out of the four non-complementary associations are binding.

Table 8.1 All possible interactions between HetB 0001, HetB* 1110, HetD 0011 and HetD* 1100

Pair	TIE (kJ/mol)
Het[BB*] 0001-1110	-159.610
Het[BD] 0001-0011	91.439
Het[B*D*] 1110-1100	169.775
Het[DD*] 0011-1100	-146.232
Het[BD*] 0001-1100	-97.910
Het[B*D] 1110-0011	-27.880

Using geometry constraints to mimic a hypothetical constrained environment is a useful first step in highlighting potentially viable letters, but it is only an approximation and neglects many important considerations, such as solvation, molecular flexibility and freedom and tautomerism.

An in-depth study (well beyond the scope of this thesis) would be required encompassing the concerns mentioned above as well as others in order to try and determine if the proposed Het or Zim alphabets could in fact be viable. This chapter will aim to explore the possible impact of molecular flexibility in terms of the geometry and construction of the particular molecules chosen to express the required D/A patterns.

In the Het set of letters a hydrogen in a terminal position is represented using an NH₂ group. If pyramidalization of this nitrogen group could occur, H-H repulsions between two planar amino groups could be lessened by the formation of a stabilizing hydrogen bond through changing one of the nitrogens from sp² to sp³ (thus H---Lp)(Fig. 8.1).

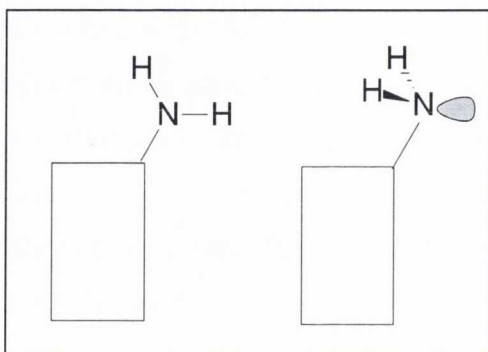


Figure 8.1 Schematic representation of nitrogen pyramidalization

8.2 Nitrogen pyramidalization

DNA can adopt several different structural conformations [1], thus it is flexible and its bases can twist and ripple. It has been shown that DNA nucleotide base molecules can adopt nonplanar structures [2-6]. The ability of nucleotides to adopt non-planar structures is required to explain the base-base interactions observed in DNA [7-11]. A specific example of nonplanarity in DNA is found in reverse Watson-Crick (RWC) pairing of GC (Fig. 5.3), in which counter rotation of the opposing amino groups occurs in order to minimise H-H repulsions [8].

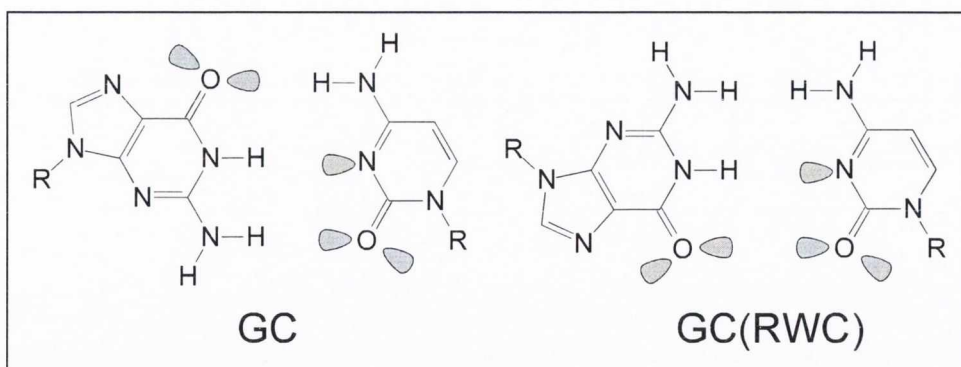


Figure 8.2 GC compared to Reverse Watson- Crick GC

In the initial study of both the ideal Het and Zim sets of letters all interactions have been confined to the C_s point group. Keeping the molecules geometrically restricted using the STRD geometry restrictions and confining all atoms strictly to the molecular plane means that any pairs which emerge as not potentially viable are truly not potentially viable. If the geometry restrictions were lessened, thus allowing the molecules a greater degree of freedom, remaining repulsions could only be reduced and not strengthened by a greater freedom of movement.

In the Het and Zim molecular sets any molecule that possesses a hydrogen in a terminal position could reduce mismatching repulsions if the amino group hydrogen is free to rotate out of the molecular plan. In order to investigate what consequence amino group pyramidalization could have on the model Het associations, studies will be carried out in this section on both the monomers and pairs which could be affected.

8.2.1 Monomer pyramidalization

In the exploration of the Het set of letters STRD conditions are used, these constrain the middle two hydrogen bonding positions (bond length and angle) in an attempt to mimic the possible constraints of a hypothetical biochemistry. As the terminal positions are free this leaves the possibility of nitrogen pyramidalization in the first and/or fourth position. 12 out of 16 Het molecules have a NH_2 group in a terminal position

[A*,B,B*,C*,D,D*,E*,F,F*,G*,H,H*](Fig. 3.3) making them potentially susceptible to pyramidalization. When pyramidalization of a nitrogen group occurs the D/A pattern of

that molecule at the point of pyramidalization is changed from a 1 to a 0. In a molecule in which pyramidalization is present this will be denoted by the presence of p after the pattern letter. In molecules in which pyramidalization is possible in one or indeed both of the terminal positions, the pyramidal position will be indicated by the Greek letter for that position. In HetA*1111, for example, a 1 is present in both of the terminal positions in theory making pyramidalization in either terminal position or indeed both potentially possible changing, one possibility as shown in Fig. 8.3 is the formation of HetA* $\rho\alpha$ 0111 (which now has the same D/A pattern as HetH 0111).

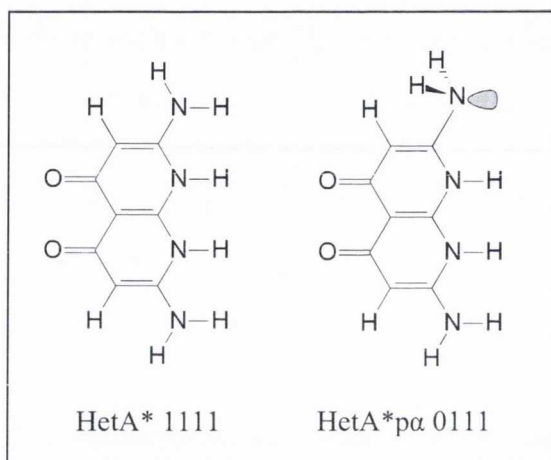


Figure 8.4 HetA*1111 compared to HetA* $\rho\alpha$ 0111

All of the Het letters in which pyramidalization is a possibility were minimised using HF with a 6-31G* basis set starting from a non-planar pyramidal structure. The energies from these nonplanar structures were then compared to the original planar geometries.

8.3 Pyramidalization-Results

8.3.1 Pyramidalized Heteronaphthalene monomers

In all cases, except for HetA*, an increase in energy is seen on comparing the standard and pyramidalized structures (Table 8.2). A spread of values for the energy change seen on pyramidalization is evident.

Table 8.2 Energy for pyramidalized structure in kJ/mol compared to standard structures.

p=Pyramidalized. Optimized using Hartree-Fock with a 6-31G(d) basis set. A negative sign in the final column implies stabilization on pyramidalization

Letter subject to Pyramidalization	Energy non-Pyramidalized. mol. X (kJ/mol)	Energy Pyramidalized mol. X _p (kJ/mol)	X-X _p (kJ/mol)
HetA* 1111 (Ap 1110)	-1772499.254	-1772505.622	-6.368
HetA* 1111 (App 0110) αδ	-1772499.254	-1772509.856	-10.602
HetB 0001 (Bp 0000)	-1576029.749	-1575957.224	72.525
HetB* 1110 (B*p 0110)	-1628115.027	-1628114.846	0.181
HetC* 1101 (C*p0101) α	-1576075.615	-1576066.734	8.881
HetC* 1101 (C*p 1100) δ	-1576075.615	-1576027.336	48.279
HetC* 1101 (C*pp 0100) αδ	-1576075.615	-1576014.648	60.967
HetD 0011 (Dp 0010)	-1772632.180	-1772615.272	16.908
HetD* 1100 (D*p 0100)	-1670160.043	-1670142.747	17.296
HetE* 1011 (E*p 0011) α	-1576075.615	-1576027.335	48.280
HetE* 1011 (E*p 1010) δ	-1576075.615	-1576066.734	8.881
HetE* 1101 (E*pp 0010) αδ	-1576075.615	-1576014.648	60.967
HetF 0101 (Fp 0100)	-1431622.083	-1431570.144	51.939
HetF* 1010 (F*p 0010)	-1431622.083	-1431570.144	51.939
HetG* 1001 (G*p 1000)	-1379506.959	-1379454.312	52.647
HetG* 1001 (G*pp 0000) αδ	-1379506.959	-1379401.532	105.427
HetH 0111 (Hp 0110)	-1628115.027	-1628114.847	0.180
HetH* 1000 (H*p 0000)	-1473557.190	-1473483.468	73.722

The spread of energies seen is not uniform across the letters theoretically susceptible to pyramidalization; some letters have significantly larger pyramidalization energy than others. If a plot of the pyramidalization energies is made (Fig. 8.5) this spread in energy values is apparent.

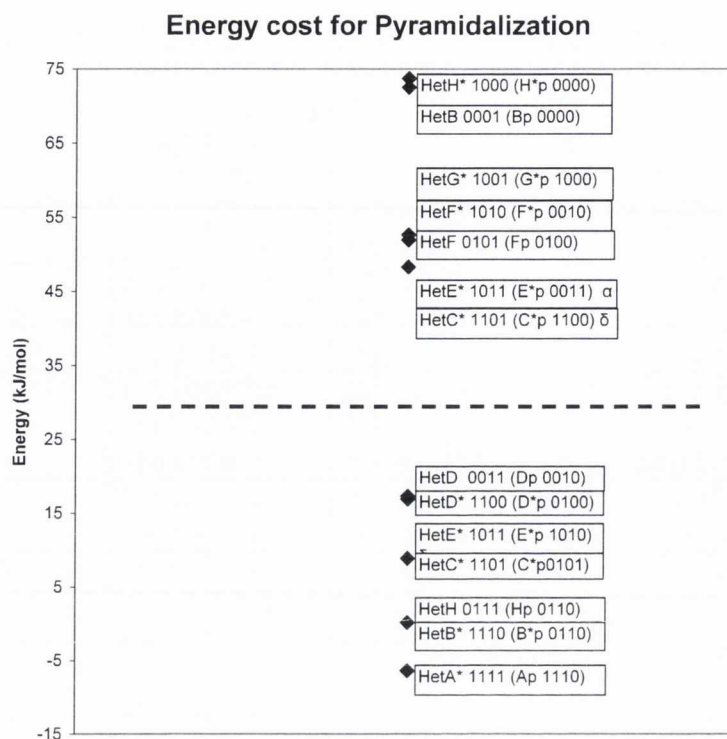


Figure 8.5 Energy cost to form pyramidalized structure in kJ/mol. Dotted line represents large gap in pyramidalization energy. This line divides the molecules into two groups.

The largest jump in energy (31 kJ/mol the position of which is marked by a dotted line in Fig 5.6) occurs between C* 1101 (C*p 1100) δ and D* 1100 (D*p 0100). This gap in the energies can be used to divide the molecules into two groups (Table 8.3).

Table 8.3 Molecules susceptible to pyramidalization broken into categories based on the cost in energy for pyramidalization

Low Energy cost for Pyramidalization	High Energy cost for Pyramidalization
A* 1111 (Ap 1110)	B 0001 (Bp 0000)
B* 1110 (B*p 0110)	C* 1101 (C*p 1100) δ
C* 1101 (C*p 0101) α	E* 1011 (E*p 0011) α
D 0011 (Dp 0010)	F 0101 (Fp 0100)
D* 1100 (D*p 0100)	F* 1010 (F*p 0010)
E* 1011 (E*p 1010) δ	G* 1001 (G*p 1000)
H 0111 (Hp 0110)	H* 1000 (H*p 0000)

In the group with a low energy difference for pyramidalization the pyramidalized NH₂ is bordered above of below by an NH (Fig. 8.6 (a)). In the remaining group of letters in which the energy cost of pyramidalization is significantly higher, the pyramidalized NH₂ is bordered in all cases by a lone pair (Fig. 8.6(b)).

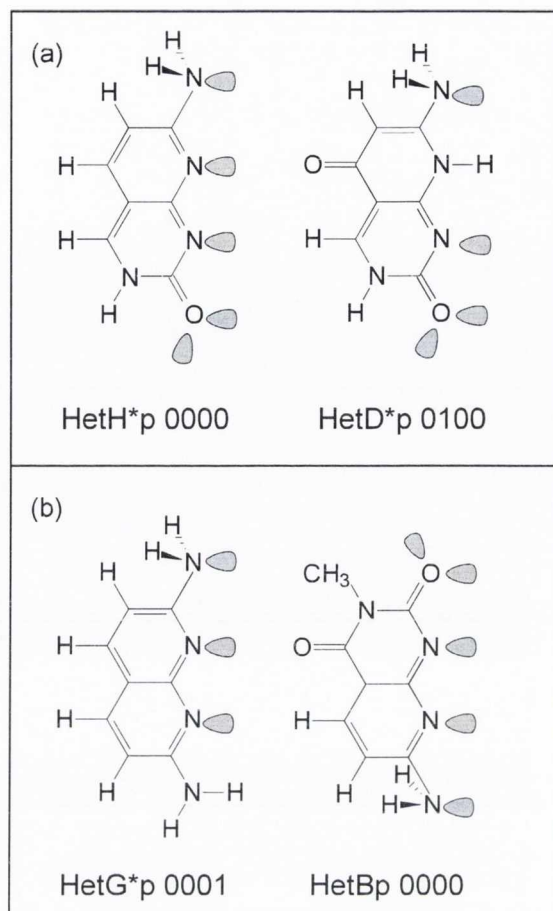


Figure 8.6 (a) HetHp(0111-0110) and HetD*p(1100-0100) show a high cost in pyramidalization
 (b) HetG*p(0101- 0100) and HetBp(1101-0101) show a low cost in pyramidalization.

The results of the exploration into pyramidalization for the Het molecules reveal that the overall change in energy depends on the D/A structure of the molecule being considered. In order to try and determine if pyramidalization even with its large energy cost at times could potentially relieve some H-H repulsions mismatching associations need to be investigated.

8.3.2 Mismatches in one position

If pyramidalization can occur it could in theory change a pair that mismatches in one or potentially two positions into a complementary pair. Het[C*F*] 1101-1010, for example, has a H-H mismatch in the alpha position, if pyramidalization of either molecule could occur this mismatch could become a match forming a pair with the pattern of FF* 0101-1010 or CC* 0010-1101 (Fig. 8.7).

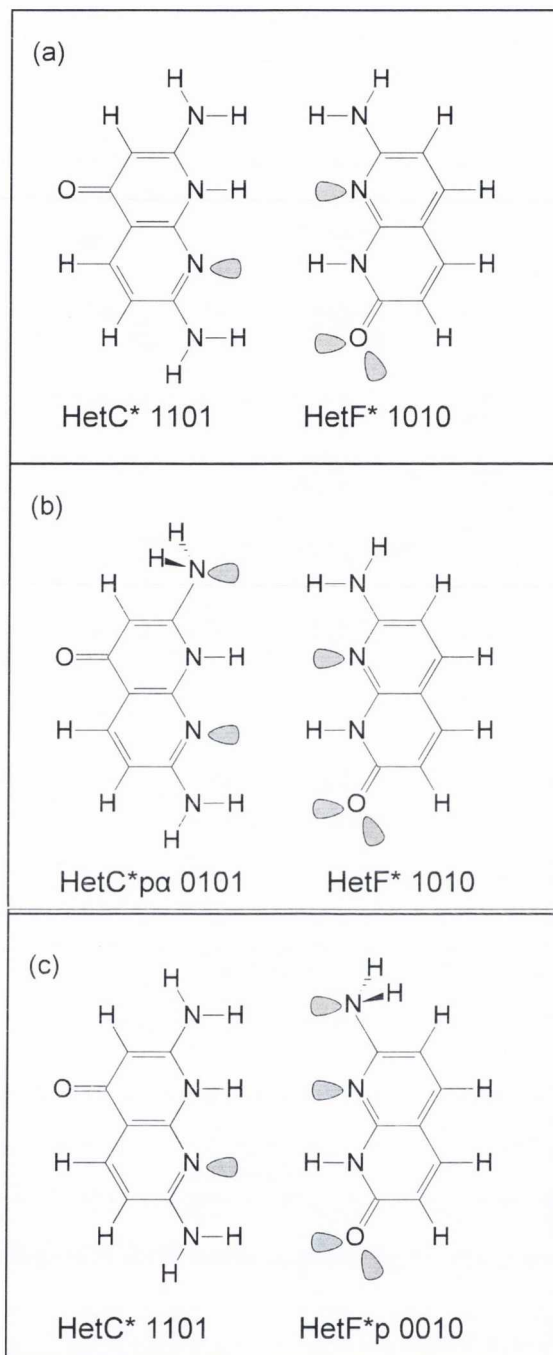


Figure 8.7 Mismatched pair (a) Het[C*F*] with one mismatch in the alpha position. On Pyramidalization of either HetC*(b) or HetF*(c) complementary pairs FF*0101-1010 (b) and CC* 0010-1101 (c) are formed.

Eight pairs that mismatch in one position could be susceptible to pyramidalization, 4 with H-H mismatches in the alpha position and 4 with a H-H mismatch in the delta position. For each of the eight pairs it is possible that pyramidalization could occur in either monomer involved in the mismatched pair implying that two calculations need to be performed for

each pair. On pyramidalization of a pair with one mismatch a D/A pattern matching one of the Het complementary associations will be formed, the specific pattern formed can be used to rationalise why the pyramidalization of one monomer results in the formation of a more binding pair than the other. All possible pyramidalization options for a sample mismatch in one position Het[F*C*] 1110-1001 can be seen below (Table 8.4). The results for this pair indicate that pyramidalization of HetF* results in the more binding association with the formation of D/A pattern CC* 0010-1101.

Table 8.4 Pyramidalization options for mismatch in one position Het[B*G*] 1110-1001 Calculated with HF 6-31G* and STRD constraints in place

X:Y	TIE(kJ/mol)	D= Difference Py-Planar	TIE Py - Cost of monomer Py
Het[F*C*] 1010-1101	9.708		
Het[F*C*ap] 1010-0101	-61.571	-71.279	-62.398
Het[F*apC*] 0010-1101	-72.300	-82.008	-30.069

All of the pairs on pyramidalization show a lower more binding TIE. On examining each of the eight associations individually, one TIE value in each set is larger than the other depending on which monomer is pyramidalized. This difference in TIE depending on which monomer is pyramidalized can be explained by examining the complementary associations that are formed on pyramidalization. As discussed in section 3.3 a range of TIE values is seen for complementary associations meaning that not all hydrogen bonds are equal in value, some combinations are in fact better than others.

Although all of the pairs which mismatch in one position that can be affected by pyramidalization are stabilized when it occurs, the overall potential viability of the Het set of letters is not altered as all mismatches in one position have already been ruled out of any potential alphabet (see section 5.2.1). This certainty that pyramidalization will not effect any potential alphabet is not the case for mismatches in two positions where it could result in the removal of letters which seem initially viable.

8.3.3 Mismatches in two positions

All of the associations in the potentially viable {Het[DD*] Het[FF*] Het[GG*]} alphabet subset (other than those that are complementary) contain two mismatches. All of these pairs which potential undergo pyramidalization were explored. The results (Table 8.5) show that on pyramidalization all of these associations could potentially become binding.

Table 8.5 TIE for mismatches between [DD*] Het[FF*] Het[GG*] including pyramidalization

X:Y	TIE (kJ/mol)	D=Difference Py-Planar	TIE Py-Cost of monomer Py
Het[DF] 0011-0101	45.968		
Het[DpF]	-17.456	-63.423	-0.548
Het[DFp]	-48.826	-94.793	3.113
Het[D*F*] 1100-1010	46.770		
Het[D*pF*]	-16.702	-63.472	0.594
Het[D*F*p]	-48.972	-95.742	2.967
Het[DG*] 0011-1001	38.960		
Het[DpG*]	-24.014	-62.974	-7.106
Het[DG*p]	-62.938	-101.898	-10.292
Het[D*G*] 1100-1001	37.624		
Het[D*pG*]	-25.070	-62.695	-7.774
Het[D*G*p]	-64.637	-102.262	-11.991
Het[FG*] 0101-1001	29.528		
Het[FpG*]	-35.230	-64.758	16.709
Het[FG*p]	-44.036	-73.564	8.610
Het[F*G*] 1010-1001	29.528		
Het[F*pG*]	-35.230	-64.758	16.709
Het[F*G*p]	-44.036	-73.564	8.610

8.4 Summary and conclusions- Pyramidalized Heteronaphthalenes

If pyramidalization can occur in the remaining letters of potential Het alphabet, non-complementary associations that bind could be formed thus breaking one of the conditions necessary for a viable alphabet. As all of the conditions for viability are no longer met when pyramidalization is taken into account, the data integrity of the potential {Het[DD*] Het[FF*] Het[GG*]} alphabet subset can not be guaranteed rendering it, most likely, non viable.

8.5 Molecular flexibility

In this work so far a set of Het molecules (Fig. 3.3) and a set based (as far as possible) on the work of Zimmerman laboratory, Zim (Fig. 7.2) have been used as molecular representations for 4-bit D/A patterns. On comparison of the two sets a broad parallel was observed suggesting that the interactions of D/A may be being captured. Both of the sets of molecules investigated so far have contained ring structures and have not been uniform in the type of hydrogen bonds present, a mixture of N-H---N and N-H---O bonds are present in varying amounts throughout the sets. To explore the effect of structural flexibility and the type of hydrogen bond present three further sets of molecules are constructed and investigated. The synthesis or stability of these molecules will not be considered here; for the purposes of this work they are used only as theoretical examples designed to give the correct flexibility and composition.

1. N-H---N heteronaphthalenes [Hetnhn]
2. N-H---N skeletal [Skelnhn]
3. Mixed Skeletal [Skelmix]

8.5.1 Adapted N-H---N Heteronaphthalenes [Hetnhn]

A set of modified Het pairs containing only N-H---N bonds was constructed in order to try and eliminate any effects caused by variation in the type of hydrogen bond present (Fig. 8.8). This set of molecules, termed Hetnhn, is based around the original Het set but instead of each complementary pair containing two N-H---N and two N-H---O hydrogen bonds (except for Het[AA*] which contains three N-H---O and one N-H---N bonds) all four are N-H---N bonds. In the Hetnhn set (and also the Skelnhn and Skelmix sets) only six pairs need to be constructed, this is possible as in the absence of a backbone structure some pairs can be repeated. Hetnhn[BB*] 0001-1110 can also be used to represent Hetnhn[HH*]0111-1000, and Hetnhn[CC*] 0010-1101 can be equal in structure to Hetnhn[EE*] 0100-1011.

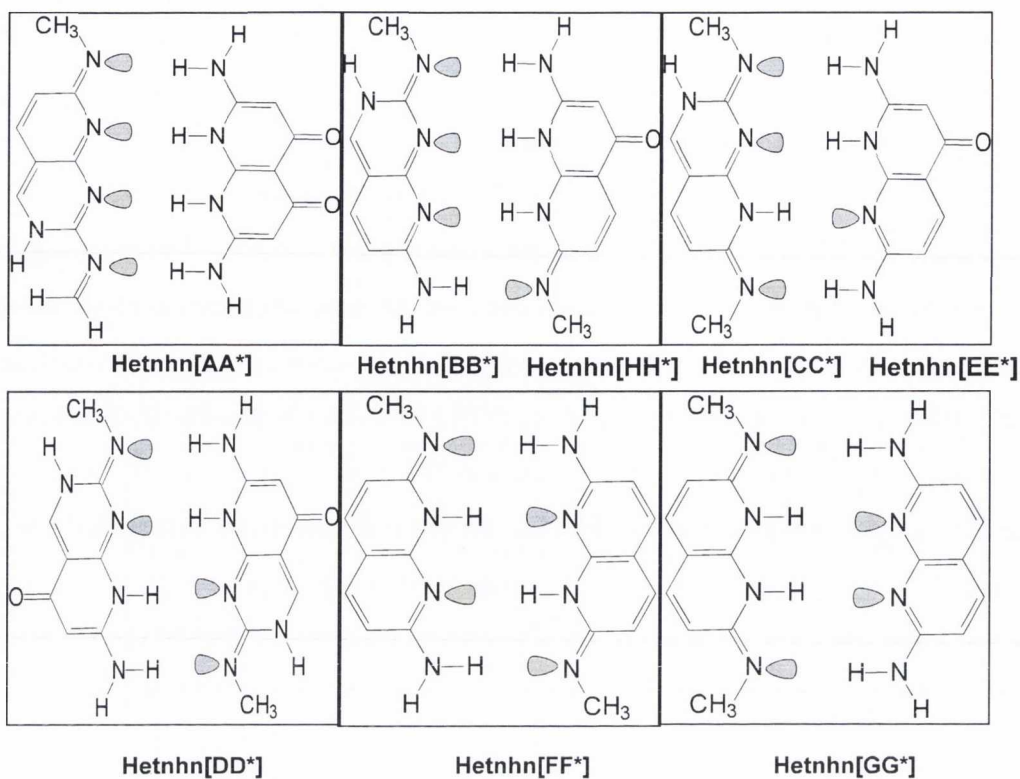


Figure 8.8 Adapted Heteronaphthalenes containing only N-H...N hydrogen bonds.

The complementary pairs were optimized using AM1 (with free geometry) and the results compared to the corresponding Het AM1 results (Fig. 8.9). Although absolute TIE values differ the same overall trend in interaction energies is evident for both data sets. In general (excluding AA* due to the oxygen as part of the ring structure in the Het set) the adapted N-H...N Heteronaphthalene pairs are less binding, on average by 15 kJ/mol. These preliminary results suggest that using a mixture of hydrogen bond types has little effect on the overall results pattern seen but does change the absolute TIE values.

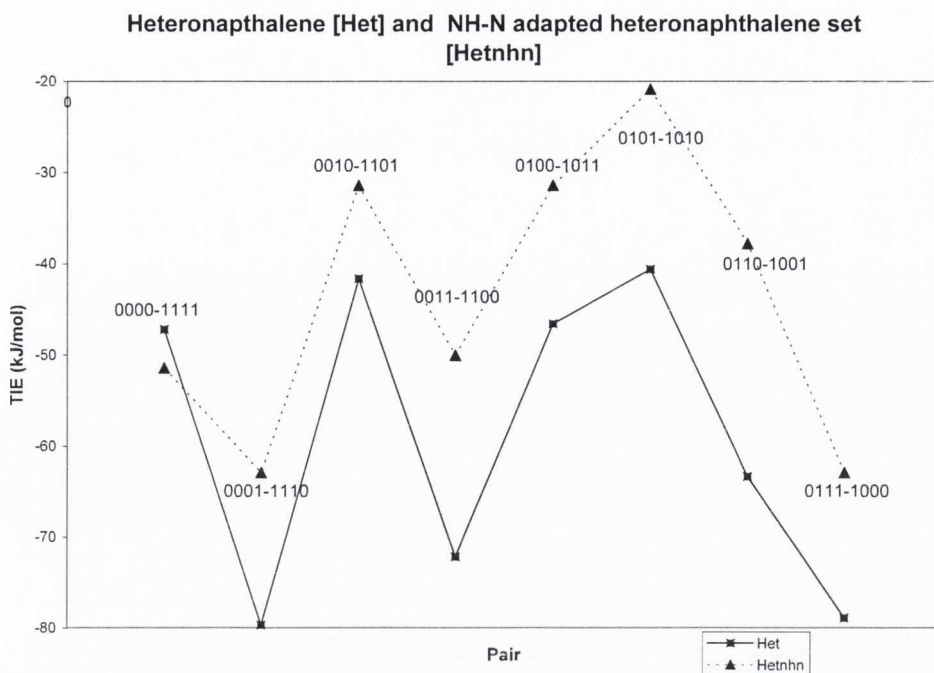


Figure 8.9 Comparison of original Het set and adapted Hetnhn set AM1.

The two sets of data can also be compared in terms of the secondary interactions present (as discussed in chapter four). The plot below (Fig. 8.10) shows the AM1 results for Hetnhn associations as well as the TIE for each predicted from the number of primary and secondary hydrogen bonds (see appendix A4 for sample calculation).

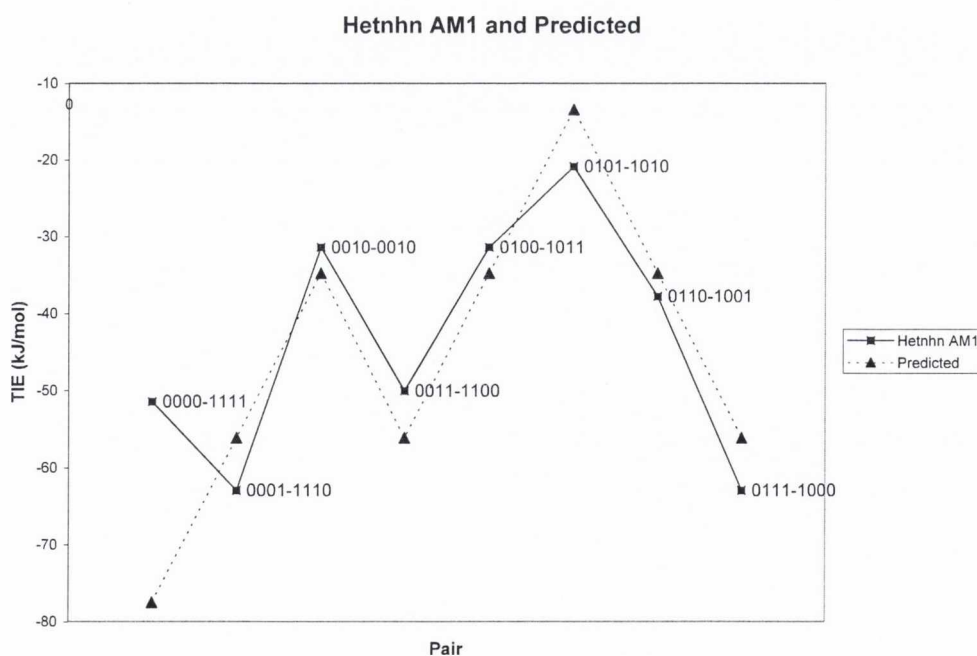


Figure 8.10 Hetnhn AM1 and TIEs predicted based on SI

This plot indicates that the predicted primary and secondary hydrogen bond values, -11.368 kJ/mol and +/- 5.340 kJ/mol respectively, provide a good overall prediction when compared to the calculated (standard deviation between the predicted and calculated data excluding AA* is 1.788). An improvement in the description of the ordering of complementary pairs through secondary interaction going from Het to the more uniform Hetnhn can be noted (Fig. 8.11). This improvement suggests that using pairs which contain more than one type of hydrogen bond can lead to deviation from the ideal as not all secondary interactions will be the same type or equal. The fact that AA* 0000-1111 is still seen to behave anomalously (even when all hydrogen bonds are equal in type) suggests that its behaviour in the Het set may only partially be due to the oxygen in the ring structure.

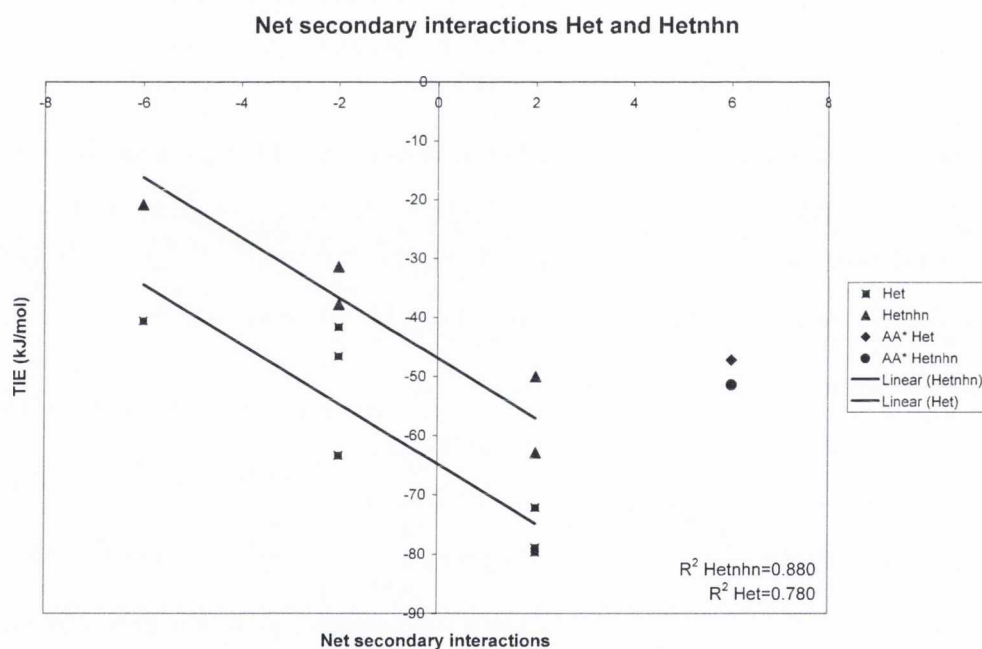


Figure 8.11 Net secondary interactions, Het and Hetnhn AM1

8.5.2 N-H---N Skeletal [Skelnhn]

This set of molecules is similar to the Hetnhn in that its pairs contain only N-H---N interactions. It has been designed in order to explore the effect of structure flexibility. In Skelnhn (Fig. 8.12) all of the molecules are chain-like in structure, allowing greater flexibility than the rigid heteronaphthalene structures previously used.

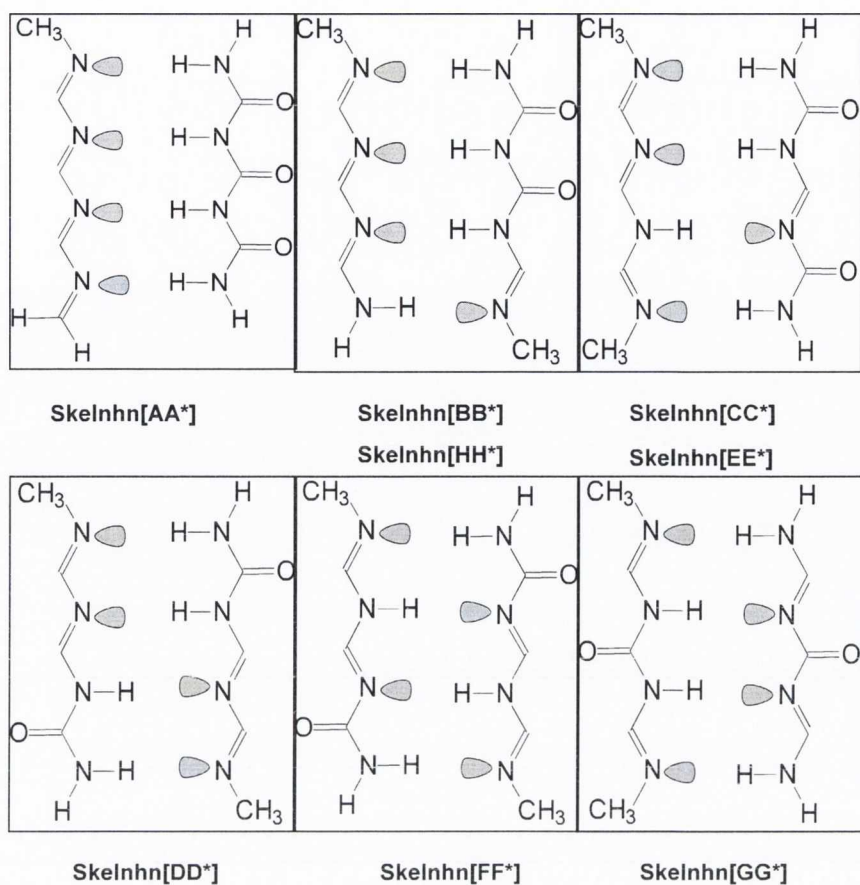


Figure 8.12 Skeletal structure containing only N-H...N hydrogen bonds.

The TIE of each of the pairs was calculated using AM1 and compared to the energies of Hetnhn (Fig. 8.13). The results indicate that a greater molecular flexibility does not in all cases lead to a lower more binding pair when compared to more rigid structures. A clear shift of energies in one direction, as was seen going from Het to Hetnhn is not evident, although both sets do follow the same overall pattern in energy values.

N-H...N Heteronaphthalene [Hetnhn] and N-H...N Skelatal [Skelnhn]

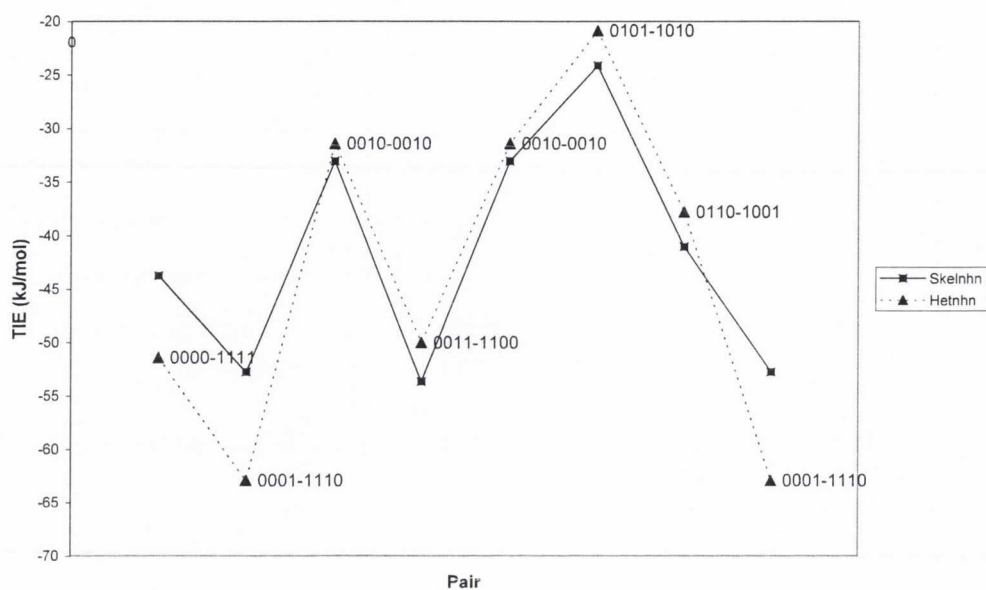


Figure 8.13 Comparison of Hetnhn set and Skelnhn AM1

Exploring the prediction of TIE based on primary and secondary interactions (- 11.024 kJ/mol and +/- 3.905 kJ/mol) (Fig. 8.14) reveals that a very close agreement can be seen (deviation 1.360 excluding AA*). An improvement in the order predicted by the number of secondary interactions can also be seen (Fig. 8.15). This suggests that with more freedom the pairs may be behaving in a more predictable fashion.

Skelnhn AM1 and Predicted

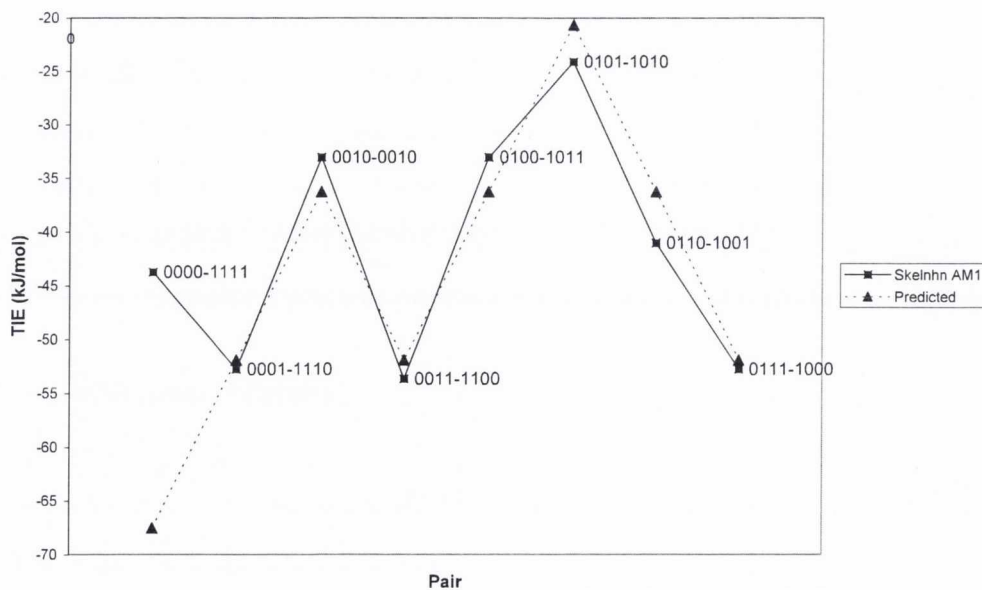


Figure 8.14 Skelnhn AM1 and TIEs predicted based on SI

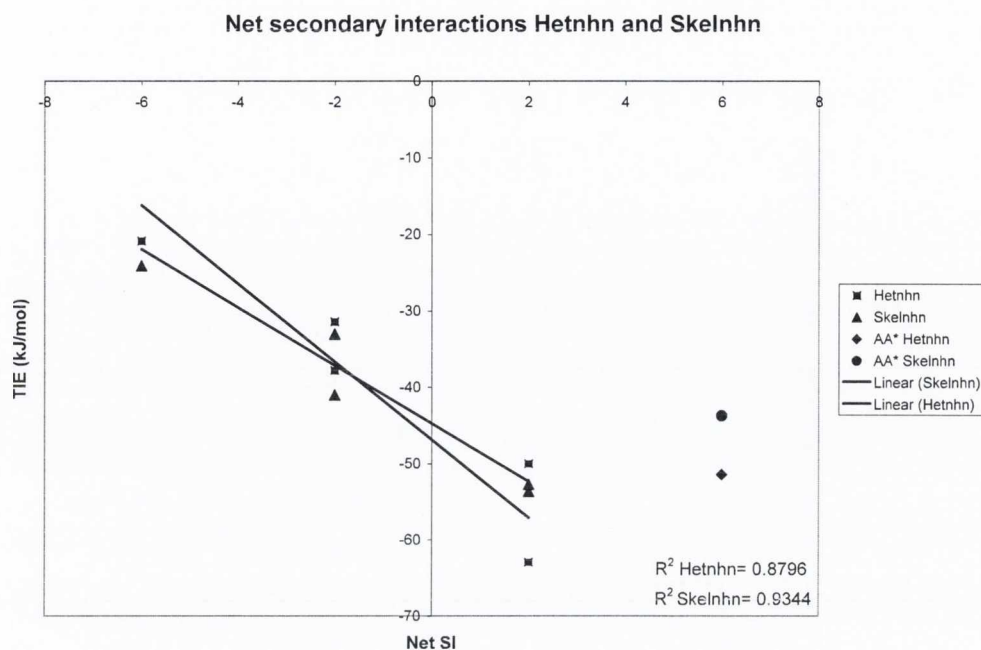


Figure 8.15 Net secondary interactions, Het and Hetnhn AM1

8.5.3 Mixed Skeletal [Skelmix]

Mixed Skeletal molecules have a simple chain-like structure similar to that of Skelnhn but they contain the same mixture of hydrogen bond types that is present in the Het set of molecules (Fig. 8.16).

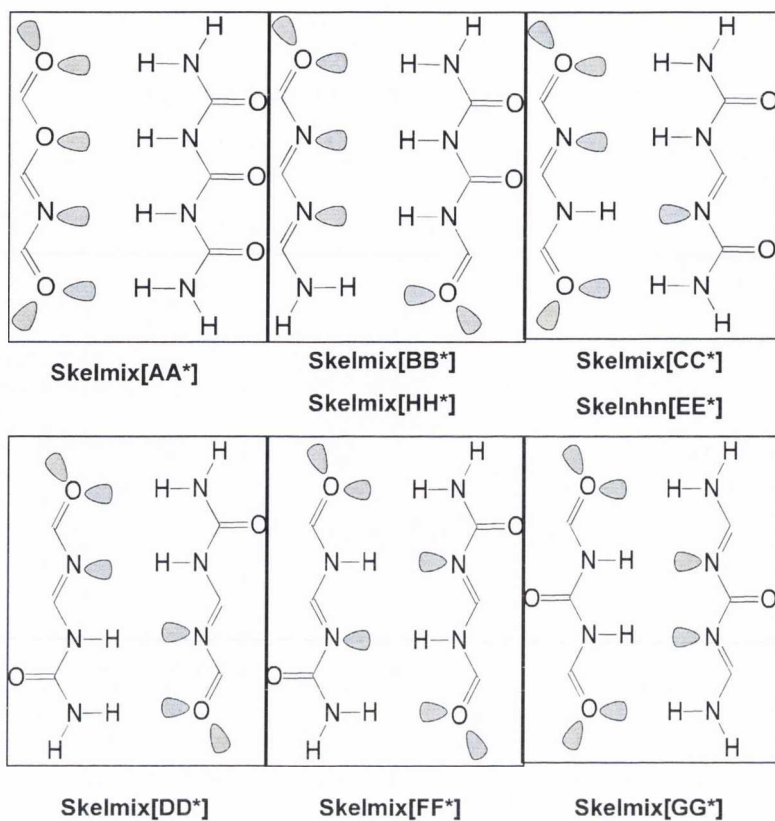


Figure 8.16 Skeletal with mixed hydrogen bonds.

Examining the interaction energies for the Skelmix pairs and comparing them to the Skelnhn (Fig. 8.17) reveals the same overall shift as was seen when comparing Het and Hetnhn. Pairs that contain only N-H---N hydrogen bonds show more repulsive energies than pairs containing a mixture of N-H---N and N-H---O bonds.

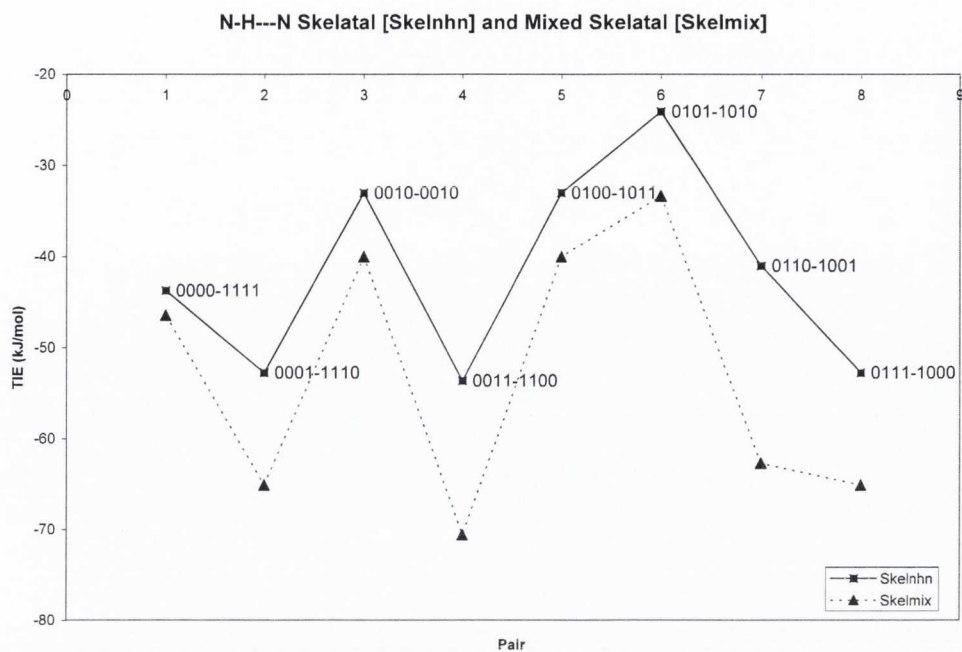


Figure 8.17 Comparison of Skelnhn set and Skelmix AM1

The interaction energies for Skelmix can also be compared to those for the Het set (Fig. 8.18) to assess the effect of flexibility. A strong agreement is seen between these two sets with several points being very close in absolute value. Interesting no case can be found in which the Skelmix set shows a lower more binding energy then the more constrained Het structures. Bringing together two molecules with a large flexibility will have a larger entropy cost then bringing together two more rigid molecules. This entropic cost could be the reason why molecules with a Skel structure do not show more binding energies then those with a less flexible structure.

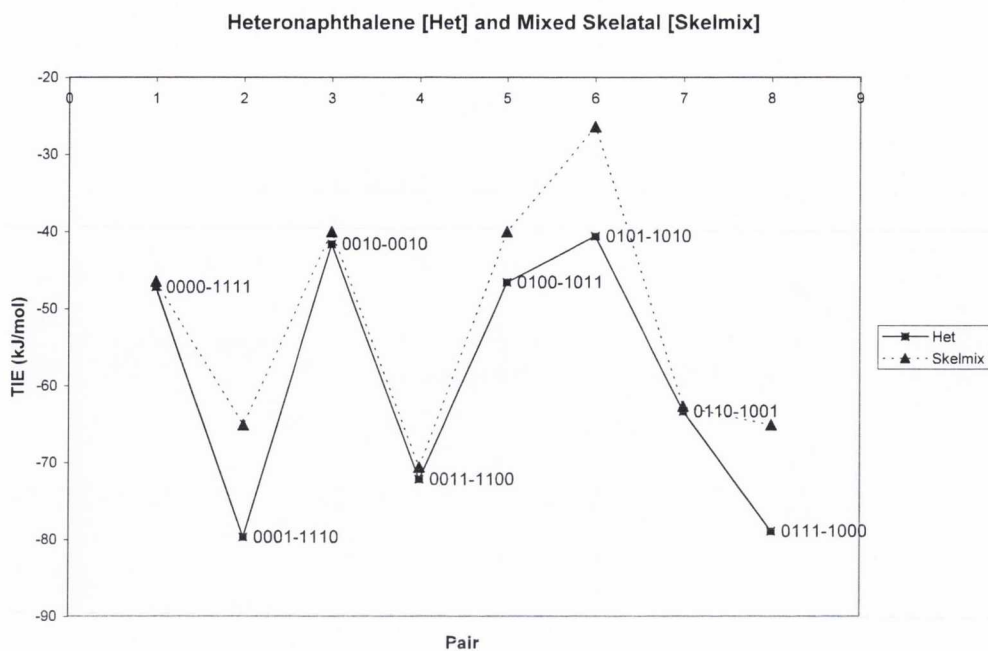


Figure 8.18 Comparison of Het and Skelmix AM1

As seen on comparing Het and Hetnbn a decrease in the ability to describe the order of TIEs by primary and secondary interactions present is seen (primary -14.339 kJ/mol secondary +/- 4.973 kJ/mol deviation excluding AA* = 4.612) (Fig. 8.19). The plot of secondary interactions (Fig. 8.20) shows a decrease in the order of result predicted by the net secondary interaction model. This result agrees with that seen on comparison of Het and Hetnbn, implying that molecules containing only one type of hydrogen bond fit the simple secondary model better.

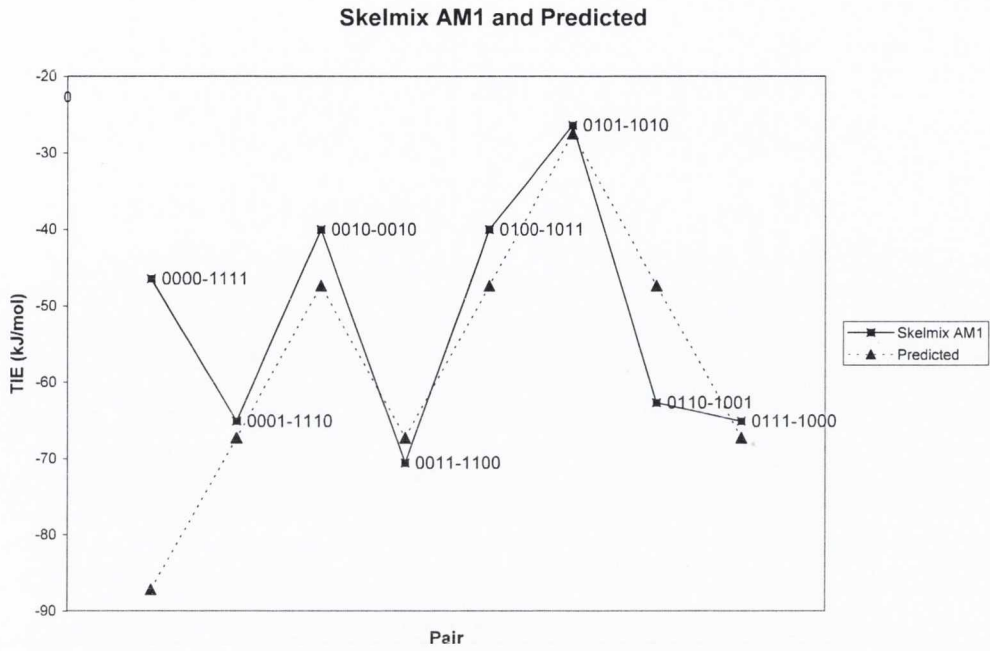


Figure 8.19 Skelmix AM1 and TIEs predicted based on SI

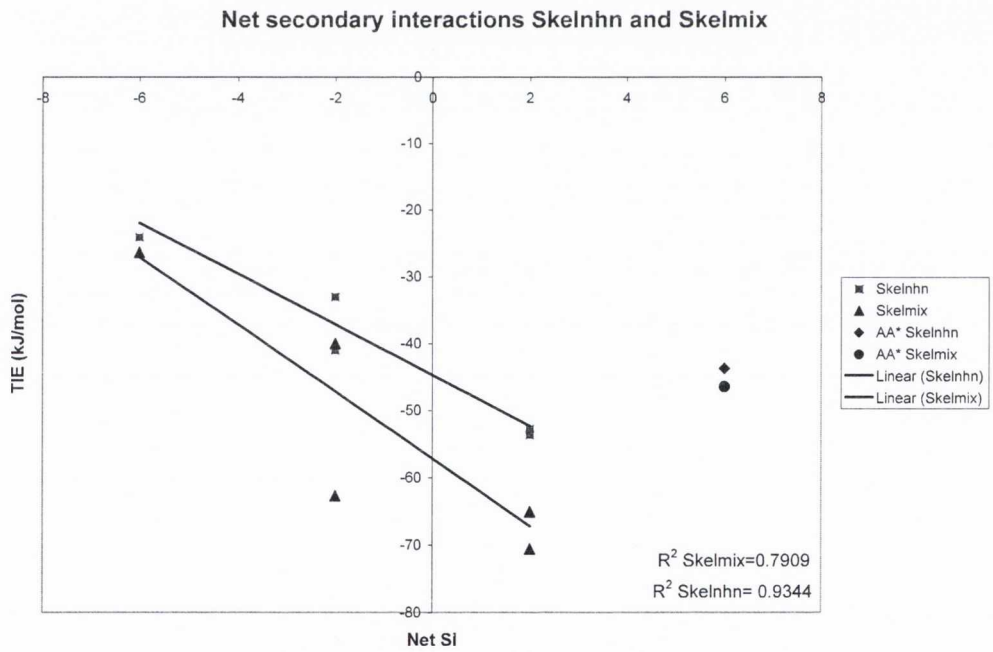


Figure 8.20 Net secondary interactions, Skelnhn and Skelmix AM1

8.6 Discussion and conclusions-All data sets

For completeness all 5 of the data sets 1 Het, 2 Zim, 3 Hetnhn, 4 Skelnhn, 5 Skelmix can be compared (Fig. 8.21)(Table 8.6).

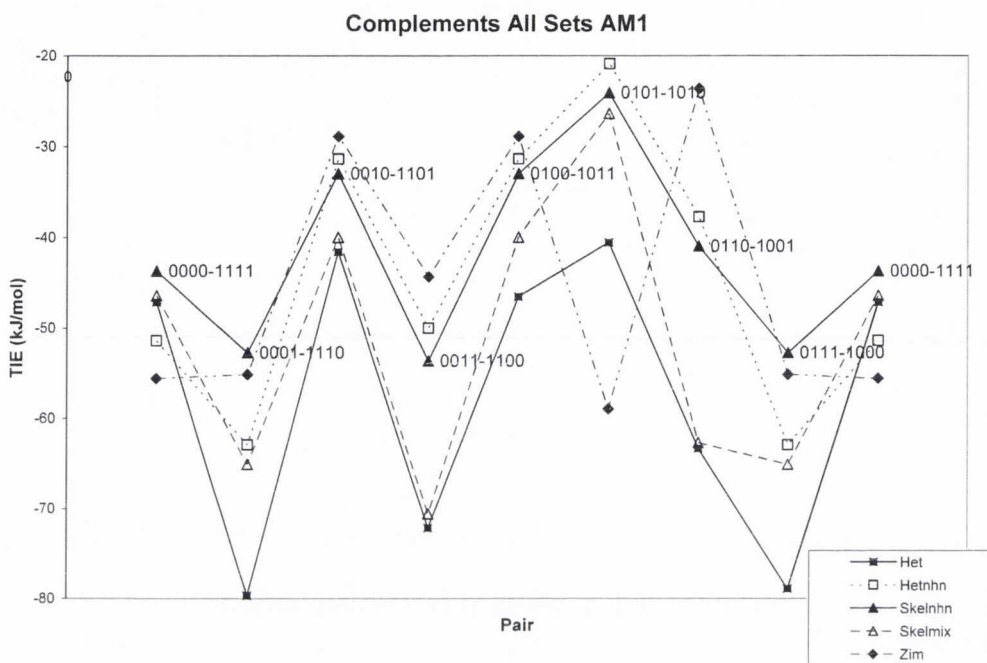


Figure 8.21 TIE all sets AM1

Table 8.6 Predicted primary and secondary hydrogen bonds all sets AM1

Set	Average Primary (AM1)	Average Secondary (AM1)	Deviation (Excluding AA*)
Het	-15.720	5.364	3.274
Hetnhn	-11.368	5.340	1.788
Skelnhn	-11.024	3.905	1.361
Skelmix	-14.339	4.973	4.612
Zim	-13.717	5.303	7.64 (Also excludes FF*)

The type of hydrogen bond used throughout a set of complements does not change the overall pattern of results as long as the type used is consistent throughout the entire set (although the Het set contains a mixture of N-H---N and N-H---O bonds each pairs consistently contains two of each apart from Het[AA*]). If sets containing different ratios of bond types are compared, the absolute TIEs will be offset as in the comparison of Skelnhn and Skelmix. Varying the ratio of bond types throughout a given set of complements will cause the resulting curve of interaction energies to be more widespread and differences between the energies calculated for different pairs to be more pronounced

as seen in the Zim set. Keeping all hydrogen bonds identical, thus making all primary and secondary interactions equal in type (as in Hetnhn and Skelnhn) does seem to improve the description of these sets using the secondary interaction prediction model. Changing the shape of complementary pairs will give approximately the same trend in energies (if the shape is constant for the set), although slight differences were noted going from a rigid ring to a more chain-like structure. The Zimmerman set of molecules examined contains both a mixture of shapes and variations in the number of hydrogen bonds of a given type present. These key differences in both structure and hydrogen bonds could account for the variation seen between the Zim and the other sets explored in this work. It is noted that the deviation in relative results seen for the Zim associations does agree with the findings of [12] who for systems of dissimilar type failed to find agreement with the secondary interaction model. AA* 0000-1111 in any of the alphabets does not follow the prediction using secondary interactions that it be the most binding of the complementary pairs. In each of the molecular sets AA* 0000-1111 shows the largest degree of wobble, causing its pair geometry to differ most from the rest of the associations. The reason why this particular pattern does not appear to have additive secondary interactions could possibly be due to a lessening of charge build up when all hydrogen bonds are aligned. The table below (Table 8.7) does show that the lone-pairs in Het[AA*] have a less negative charge when compared to other pairs.

Table 8.7 Mulliken charges [13] for sample Het associations

Mulliken charges			
Het[AA*]			
HetA 0000			HetA* 1111
O	-0.283	0.237	H
O	-0.179	0.249	H
N	-0.256	0.255	H
O	-0.312	0.238	H
Het[BB*]			
HetB 0001			HetB* 1110
O	-0.374	0.262	H
N	-0.307	0.289	H
N	-0.310	0.275	H
H	0.289	-0.431	O
Het[DD*]			
HetD 0011			HetD* 1100
O	-0.394	0.272	H
N	-0.358	0.302	H
H	0.300	-0.357	N
H	0.271	-0.390	O

Several general preliminary conclusions can be drawn as summarized in Table 8.8;

Table 8.8 summary of main shape and flexibility findings

1	Molecular pairs containing only N-H--N bonds show higher less attractive TIEs than pairs containing a mixture of N-H---N and N-H---O bonds.
2	Sets of molecules, such as Zim, which are inconsistent in the flexibility each molecule has and vary in the type and ratio of hydrogen bonds present show larger more pronounced differences between pairs.
3	The more flexible and more uniform in terms of the type of hydrogen bond present throughout a data set the easier it is to predict the ordering of complementary associations using the secondary interactions present.

1. Dickerson, R.E., et al., *The anatomy of A-DNA, B-DNA, AND Z-DNA*. Science, 1982. **216**(4545): p. 475-485.
2. Riggs, N.V., *An Ab Initio study of the stationary structures of the major gas phase tautomer of adenine*. Chemical Physics Letters, 1991. **177**(4-5): p. 447-450.
3. Leszczynski, J. *Are the amino groups in the nucleic acid bases coplanar with the molecular rings Ab Initio HF 6-31G* and MP2 6-31G* studies. 32nd Sanibel International Symp on the Application of Fundamental Theory to Problems of Biology and Pharmacology*. 1992. St Augustine, Fl: John Wiley & Sons Inc.
4. Sponer, J. and P. Hobza, *Nonplanar geometry of DNA bases- Ab Initio 2nd order Moller-Plesset study*. Journal of Physical Chemistry, 1994. **98**(12): p. 3161-3164.
5. Dong, F. and R.E. Miller, *Vibrational transition moment angles in isolated biomolecules: A structural tool*. Science, 2002. **298**(5596): p. 1227-1230.
6. Wang, S.Y. and H.F. Schaefer, *The small planarization barriers for the amino group in the nucleic acid bases*. Journal of Chemical Physics, 2006. **124**(4): p. 8.
7. Sponer, J. and P. Hobza, *Bifurcated hydrogen bonds in DNA crystal structures- An ab initio quantum chemical study*. Journal of the American Chemical Society, 1994. **116**(2): p. 709-714.
8. Sponer, J. and P. Hobza, *DNA base amino groups and their role in molecular interactions: Ab initio and preliminary density functional theory calculations*. International Journal of Quantum Chemistry, 1996. **57**(5): p. 959-970.
9. Luisi, B., et al., *On the potential role of the amino nitrogen atom as a hydrogen bond acceptor in macromolecules*. Journal of Molecular Biology, 1998. **279**(5): p. 1123-1136.
10. Hobza, P. and J. Sponer, *Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations*. Chemical Reviews, 1999. **99**(11): p. 3247-3276.
11. Li, X.Q. and P. Fan, *A duplex DNA model with regular inter-base-pair hydrogen bonds*. Journal of Theoretical Biology, 2010. **266**(3): p. 374-379.
12. Popelier, P.L.A. and L. Joubert, *The elusive atomic rationale for DNA base pair stability*. Journal of the American Chemical Society, 2002. **124**(29): p. 8725-8729.
13. Leach, A.R., *Molecular Modelling Principles and Applications (second edition)*. Prentice Hall, 2001

9 Conclusions

9.1 Introduction

One of the most fundamental questions in molecular biology is why nature has chosen A, C, G, U/T for the genetic alphabet. It has been shown [1] that some insight into nature's choice can be gained by assigning each letter (molecule) a numerical representation based on the hydrogen and lone pair pattern of each molecule. This numerical representation approach can be taken further, as in this thesis, and used to explore other sets of letters to determine if perhaps they could form a potentially viable molecular alphabet.

A vast number of molecules varying widely in shape and size are capable of molecular recognition. Considering the large number of possibilities, why has nature chosen to use nucleotides; could an alphabet potentially be composed of something different? In order to explore this, a set of Het molecules was designed and constructed each with 4 hydrogen D/A positions, differing from nucleotides in that all bits of each pattern come directly from the D/A motif (Fig. 9.1).

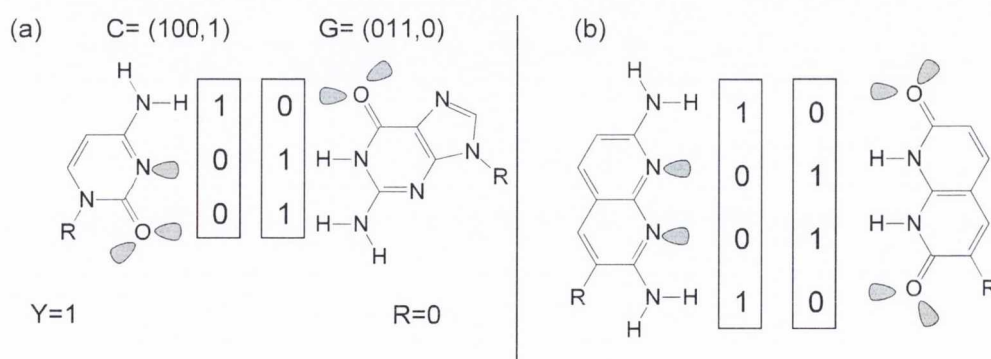


Figure 9.1 (a)Nucleotide pair CG (b)Informationally equivalent heteronaphthalene pair

For the overall concise conclusions of this work and its context see the summary at the start of the thesis.

9.2 A Heteronaphthalene alphabet

In order to model the Het potential alphabet set of letters (Fig. 3.3) appropriate geometry constraints were determined and used in the absence of an enzyme environment. The TIE

of all possible combinations of pairs was explored using a variety of computational methods (HF 6-31G* basis set, with and without BSSE correction, AM1, PM3 and MP2).

The results for the complementary associations revealed a spread of interaction energy values indicating that not all arrays with equal numbers of hydrogen bonds give similar energies. Some hydrogen bonding arrays are stronger than others. The difference in TIE between the 8 complementary pairs was explored by taking secondary interactions into account. A simple equation was put in place that could potentially explain the ordering of the complementary Het TIE based on the secondary interactions present and in the position in which these SI occur.

In order to determine if the proposed potential Het alphabet or a subset of it could possibly be viable, all TIE calculated were analysed using (in the first instance) the first two conditions for viability (Table 9.1).

Table 9.1 Table of viability requirements

1	Each molecule must bind to its complementary molecule
2	Each molecule should repel any molecule with which it does not form a complementary pair.
3	Any surviving molecules must comply with chemical constraints (although this will not be a primary concern in this thesis).

In the Het potential alphabet letter set three pairs were found that comply with the first two conditions for viability: Het[DD*], Het[FF*], and Het[GG*](Fig. 9.3). For these three pairs all combinations of letters which result in a mismatch being present have an energy sufficiently repulsive to prevent binding. This solution differs from that seen for the nucleotide alphabet, where based on patterns alone (before the consideration of chemical constraints) 8 letters were potentially viable.

As was the case for complementary associations, all pairs with the same number of mismatches or even mismatches in the same positions did vary in energy values. A fit

could not be designed for secondary interactions in mismatching associations as many of the pairs have equal or no net secondary interactions.

Table 9.2 All possible pairings between Het[DD*], Het[FF*], Het[GG*]

Pair	TIE (kJ/mol)
XNOR 0011	
Het[FG*] 0101-1001	29.528
Het[F*G] 1010-0110	52.188
XNOR 1100	
Het[F*G*] 1010-1001	29.528
Het[FG] 0101-0110	52.187
XNOR 0101	
Het[DG*] 0011-1001	38.96
Het[D*G] 1100-0110	41.572
XNOR 1010	
Het[D*G*] 1100-1001	37.624
Het[DG] 0011-0110	40.431
XNOR 0110	
Het[D*F] 1100-0101	34.334
Het[DF*] 0011-1010	34.741
XNOR 1001	
Het[DF] 0011-0101	45.968
Het[D*F*] 1100-1010	46.77
XNOR 0000	
Het[DD*] 0011-1100	-146.232
Het[FF*] 0101-1010	-86.551
Het[GG*] 0110-1001	-124.236

BSSE was taken into account and removed from all of the TIE results. Whilst the removal of BSSE did make each association more repulsive its effect was seen to be almost constant across the entire set. Although the absolute TIE values were changed when BSSE was removed the overall relative pattern of energies remained unchanged.

It is important to determine if the results achieved are truly representative of the Het set of molecules and whether they are independent of any particular computational method. Semi-empirical methods AM1, PM3 and MP2 were used to verify the results for the Het letters. All three methods confirmed the same overall result, only three pairs can possibly coexist Het[DD*], Het[FF*], and Het[GG*].

9.3 A Zimmerman alphabet

It is as important as the use of different calculation methods as discussed above to establish if the results determined for the ideal Het letters are linked to the 4-bit D/A patterns rather than only to the molecules used to represent them. To investigate if this is in fact the case an alternative set of molecules from which a potential alphabet could be formed was explored. This alternative set was based as far as possible on molecules studied in the work of the Zimmerman laboratory. This new set of molecules differs from the ideal Het set in that they are no longer all uniform in shape and structure. All possible Zim associations were explored in the same way as the Het and the results of the sets compared. The overall result seen for the Zim letters was similar to that of the Het. Two potentially viable subsets each containing four letters emerge (Table 9.3a, Table 9.3b).

Table 9.3a All possible pairings between Zim[DD*], Zim[GG*]

See chapter 7 for Zim[GG*] TIE explanation

XNOR 0101	TIE kJ/mol
Zim[DG] 0011-0110	47.453
Zim[D*G*] 1100-1001	93.355
zim[DG*] 0011-1001	93.355
Zim[D*G] 1100-0110	47.453
Zim[DD*] 0011-1100	-93.833
Zim[GG*] 0110-1001	4.312

Table 9.3b All possible pairings between Zim[DD*], Zim[FF*]

See chapter 7 for Zim[GG*] TIE explanation

XNOR 0011	TIE kJ/mol
Zim[FG] 0101-0110	71.948
Zim[F*G*] 1010-1001	108.352
Zim[FG*] 0101-1001	108.352
Zim[FG] 1010-0110	71.948
Zim[GG*] 0110-1001	4.312
Zim[FF*] 0101-1010	-87.145

Differences in the absolute TIE values are evident between the two alphabets. This is to be expected as the molecular sets differ in structure from each other but even within the Zimmerman set a wide deviation in shape and overall rigidity is apparent. Zimmerman and Corbin [2] have also reported differences in association constants for arrays with the same pattern differing in structure. BSSE was not removed from the Zimmerman alphabet as

doing so would simply shift the TIEs to higher energies, but as they would all be moved by a similar amount the total result will remain unchanged (Fig. 9.2).

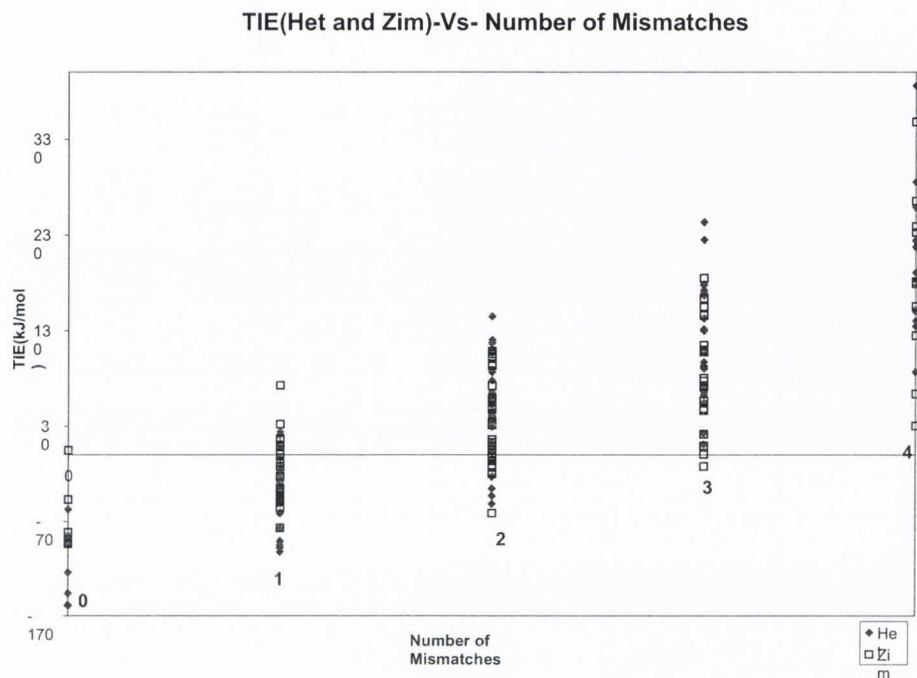


Figure 9.2 TIE for versus number of mismatches for Zim and Het associations

The results for the Het and Zim alphabets suggest that a small subset of between 4 and 6 letters could possibly coexist at the maximum mutual distance for an alphabet with 4 D/A positions ($\delta = n/2$ $n=4$) as defined by Mac Dónaill [3] (see section 1.4). If something such as size was used to divide the Het and Zim alphabets into groups in such a way that one molecule in each complementary association would belong to each group, a significant advantage and possible widening of the number of potentially viable letters would be created. If the associations were partitioned in a way such a size it would allow the minimum distance in a potentially viable group of letters to increase from $\delta = 2$ to $\delta = 3$ thus decreasing the chance of a non-complementary association with a binding TIE occurring.

9.4 Pyramidalization

Any potentially viable set of letters in addition to meeting the necessary energy requirements must also meet with possible chemical limitations. One such consideration is the possibility of pyramidalization of terminal NH₂ groups: if these groups could move out of the molecular plane, interactions between mismatching hydrogen could be greatly lessened and if full pyramidalization occurs a match could be created instead of a mismatch (Fig. 9.3). If this phenomenon did occur it could lead to potentially viable associations becoming non viable.

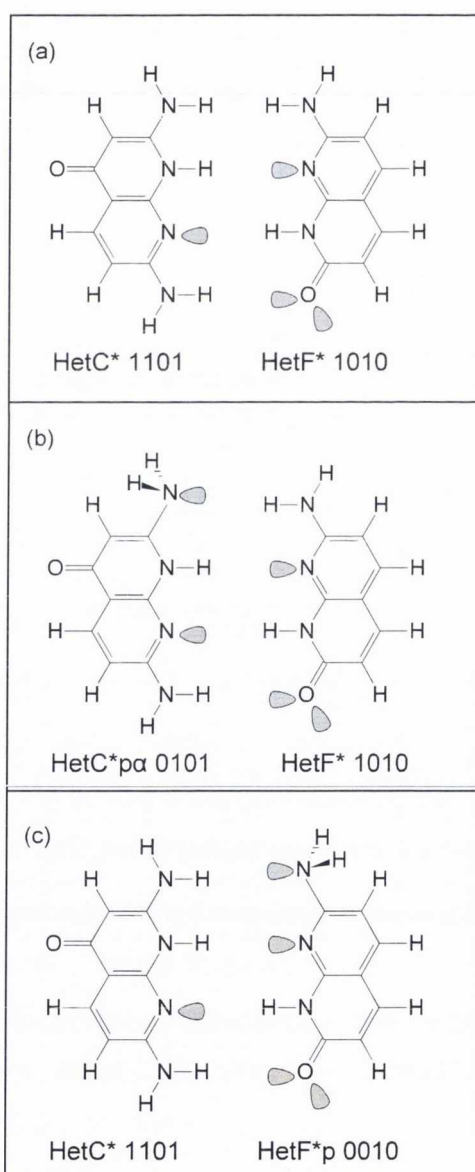


Figure 9.3 Mismatched pair (a) Het[C*F*] with one mismatch in the alpha position. On Pyramidalization of either HetC*(b) or HetF*(c) complementary pairs FF*0101-1010 (b) and CC* 0010-1101 (c) are formed.

If the potentially viable alphabets considered here were developed further, attached to a DNA analogous backbone structure, and given a greater freedom of movement which enable pyramidalization to some degree, the potentially viable Het and Zim sub-sets containing a mixture of DD*, FF* and GG* all with terminal hydrogens could become non-viable.

9.5 Molecular flexibility

To attempt to gain some insight into the role that the molecular flexibility of a potential alphabet could have on the variation and uniformity of results seen, several other alphabet sets were designed and the complementary associations for each explored at the semi-empirical level. In total 5 potential sets of 4 D/A position molecules were considered;

1. Heteronaphthalenes
2. Zimmerman
3. Linear with the heteronaphthalene patterns
4. NH-N only rings
5. NH-N only linear

The 8 complementary associations for each set of molecules did show the same overall trend in results but not identical TIE values. The shape and structure of the molecular set chosen does affect the uniformity in behaviour seen between sets and also within a given set. The Zim alphabet has the largest deviation in relative result pattern compared to the others sets. This is most likely the case because due to effects such as large steric interactions caused by particular structural choices (Fig. 9.4).

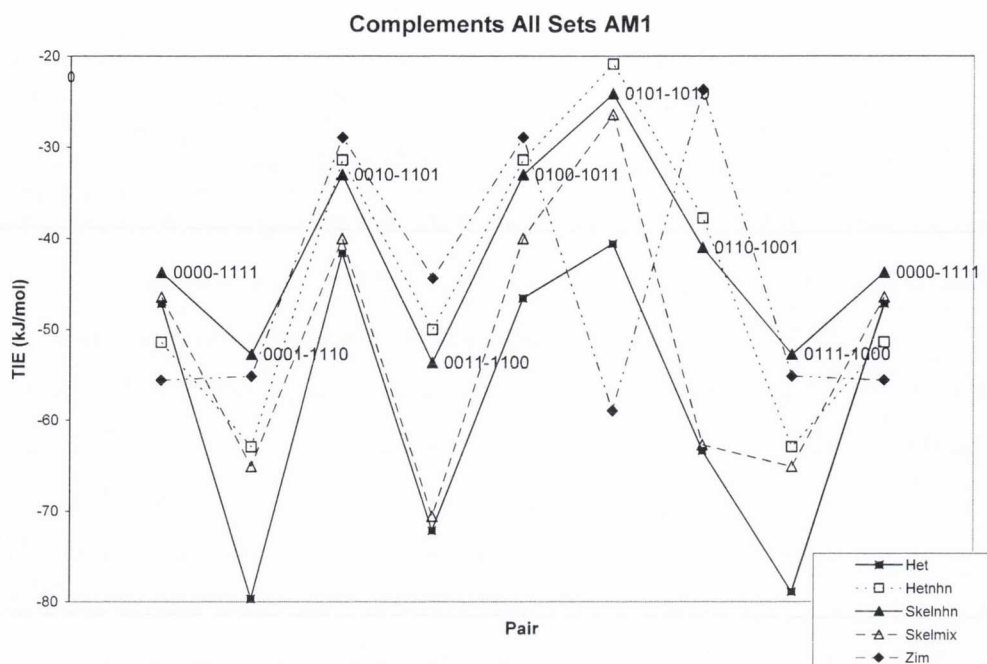


Figure 9.4 Complementary associations AM1 free for all data sets

9.6 Future work

Although not the primary concern of this thesis, all possible chemical limitations such as tautomerism or susceptibility to hydrolysis would need to be considered, in order to determine if any of the initially viable subsets of letters is truly viable. Further exploration could also be carried out to complete the study of the Hetnhn, Skelnhn and Skelmix sets of letters. Comparing the behaviour of mismatching associations across all data sets could provide a better understanding of how TIE changes with the number of mismatches. This would also be a useful way to determine if a viable subset of letters between DD*, FF* and GG* does exist. Due to the anomalous behaviour of Zim[GG*] and Zim[FF*] it is difficult to say with certainty if all three pairs can coexist.

Since the completion of the experimental work presented in this thesis Blight et al. [4] have synthesised a stable AAAA-DDDD array. This could be used to test the anomalous behaviour noted for AA* in all molecular sets.

In order to further the study of alternatives to the nucleotide alphabet several avenues could be explored.

- The effects of solvation
- The addition of a backbone anchoring structure
- A larger R group being attached in a uniform matter throughout a letter set
- Relaxation of geometry restrictions
- Set of letters with more than 4 D/A positions
- Further explorations into molecules with equal D/A arrays but different structures
- Alphabets with more than 4 D/A positions

1. Mac Donaill, D.A., *Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition*. *Origins of Life and Evolution of the Biosphere*, 2003. **33**(4-5): p. 433-455.
2. Zimmerman, S.C. and F.S. Corbin, *Heteroaromatic modules for self-assembly using multiple hydrogen bonds*. *Molecular Self-Assembly*, 2000. **96**: p. 63-94.
3. Mac Donaill, D.A., *Molecular Error-Coding: Why Nucleotides Come in Two Sizes* In preparation.
4. Blight, B.A., et al., *An AAAA-DDDD quadruple hydrogen-bond array*, *Nature Chemistry*, 2011. **3**(3): p. 244-248.

Appendices

A1 Hydrogen bonds

The hydrogen bond is important in all areas of science. It is a vital part of DNA as seen in section 1.1.1. Hydrogen bonds are responsible for holding base pairs together. A recent report by IUPAC proposed a definition of the hydrogen bond.

The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule or fragment in which there is evidence of bond formation.[1]

A typical hydrogen bond may be denoted as X-H ••• Y-Z. In this notation X-H is the referred to as the bond donor and Y-Z the acceptor. The three dots are used to represent the hydrogen bond itself. Classically only electronegative elements such as nitrogen, fluorine and oxygen were thought to be involved in hydrogen bonding but over time new elements have been added to this. The current IUPAC definition (above) states that in fact any element that is more electronegative than hydrogen may be involved in a hydrogen bond.

Hydrogen bonds are often broken into three categories: weak, medium and strong. These categories are defined by the bond distance and energy. In the report giving a data range for hydrogen bond distances (heavy atom to heavy atom or otherwise) and also specific energy values is avoided. In respect to hydrogen bond distances the report comes down strongly against using the van der Waals radii as a measure of the hydrogen bond distance. It does however list data as given by Jeffery, Desiraju and Steiner as giving “reasonable” values for heavy atom to heavy atom distances. Sample results taken from Desiraju and Steiner ([2] pg. 13) and Jeffrey ([3] pg. 65) can be seen below (Table A1.1, Fig. A1.1);

Table A1.1 Hydrogen bond energies and distances

	Very strong	Strong	Weak
Bond energy (kJ/mol)	63-167	17-63	<17
Example	P-OH•••O=P	O-H•••O-H	C-H•••O
Bond Distance (Heavy atom) Å	2.2-2.5	2.5-3.2	3.0-4.0

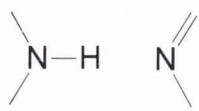
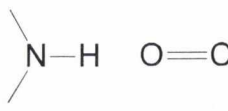
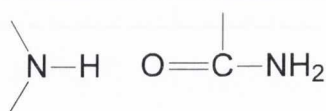
 <p>Minimum 1.73Å Maximum 2.23Å</p>	 <p>Minimum 1.69Å Maximum 2.32Å</p>
 <p>1.93 Av Å</p>	

Figure A1.1 Hydrogen bond distances ranges. Distances here are measured from hydrogen to acceptor.

These ranges will be used as a rough guide to clarify the results seen in this thesis.

In exploring the energy of hydrogen bonds the report concludes that “It is clear that specifying an energy cut-off is arbitrary and does not help in identifying or excluding the possibility of a hydrogen bond being present.” ([1], pg. 8). Bearing this in mind the energies given above for the different bond categories will only be considered loosely.

A1.1 bond lengths for complementary Heteronaphthalene associations

Comparing the hydrogen bond lengths determined for complementary pairs in the Het set to those in our genetic alphabet, the bond lengths for nucleotide base pairs, A:T and C:G (bond lengths as cited in [4]) (Fig. A1.2) are shorter than those calculated for Het pairs.

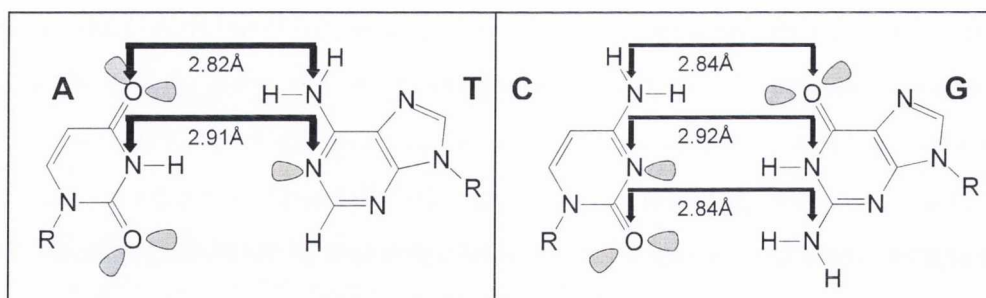


Figure A1.2 Hydrogen bond lengths for base pairs AT and CG.

In comparing the bond lengths of the two alphabets it should be remembered that those seen for the Het pairs are purely theoretical (whilst those quoted for the genetic alphabet are experimentally determined) and will depend on the particular computational method and basis set chosen. The base pairs in DNA also possess a greater degree of freedom and can twist and bend out of the plane. Although the results of the Het associations are longer

in comparison to nucleotides they are not outside the normal region. Values measure from hydrogen to acceptor (rather than heavy atom to heavy atom) show a range of 1.85 Å – 2.58 Å for the Het complementary pairs.

1. Arunan, E., et al., *Definition of a Hydrogen Bond*. IUPAC task group, 2010.
2. Desiraju, G.R., Steiner, T, *The Weak Hydrogen Bond, In Structural Chemistry and Biology*. 1999: OUP.
3. Jeffrey, G.A., *An Introduction to Hydrogen Bonding*. 1997: OUP.
4. Sinden, R.R., *DNA structure and function*. 1994: Academic Press

A2 BSSE calculation

The result of calculations following each of the three paths can be seen in (Table A2.1). For diagram of molecular pairs see Fig. 3.3 and for bond position labelling see Fig. 3.1.

Table A2.1 Dimer energy and TIE for each of the Het complementary pairs, determined using different BSSE correction paths. HF 6-31G(d) basis set

Pair	TIE (kJ/mol)	BSSE (kJ/mol)
Het[AA*] 0000-1111		
Path 1 [OPT No CP]	-73.522	
Path 2 [OPT then CP]	-64.743	8.779
Path 3 [OPT with CP]	-64.824	8.698
Het[BB*] 0001-1110		
Path 1 [OPT No CP]	-160.883	
Path 2 [OPT then CP]	-145.778	15.105
Path 3 [OPT with CP]	-145.812	15.070
Het[CC*] 0010-1101		
Path 1 [OPT No CP]	-88.961	
Path 2 [OPT then CP]	-74.732	14.229
Path 3 [OPT with CP]	-74.868	14.093
Het[DD*] 0011-1100		
Path 1 [OPT No CP]	-147.449	
Path 2 [OPT then CP]	-133.536	13.913
Path 3 [OPT with CP]	-133.285	14.164
Het[EE*] 0100-1011		
Path 1 [OPT No CP]	-92.783	
Path 2 [OPT then CP]	-78.148	14.635
Path 3 [OPT with CP]	-78.305	14.478
Het[FF*] 0101-1010		
Path 1 [OPT No CP]	-88.165	
Path 2 [OPT then CP]	-71.766	16.399
Path 3 [OPT with CP]	-71.906	16.260
Het[GG*] 0110-1001		
Path 1 [OPT No CP]	-124.781	
Path 2 [OPT then CP]	-108.021	16.760
Path 3 [OPT with CP]	-108.135	16.646
Het[HH*] 0111-1000		
Path 1 [OPT No CP]	-159.981	
Path 2 [OPT then CP]	-144.777	15.204
Path 3 [OPT with CP]	-144.916	15.065

Very little difference in TIE is observed between path 2 and 3, indicating that BSSE in the case of these molecular pairs does not vary greatly (on average 0.1kJ/mol) depending on the superposition error calculation path taken. The average BSSE across the 8 pairs is 14.378 kJ/mol (path 2) and 14.309 kJ/mol (path 3). The standard deviation (appendix A6) was determined to be 2.308 (path 2) and 2.291 (path 3). There is only a very small change

in the BSSE value whichever calculation path is taken. The principal difference between paths 2 and 3 is the dimer geometry, in path 3 BSSE is calculated during the optimisation procedure and altered accordingly during the calculation. The four hydrogen bond positions were studied for each of the complementary pairs for the different BSSE exclusion method paths available (Table A2.2).

Table A2.2 Hydrogen bond distances for the complementary pairs using different BSSE removal paths

Geom. Analysis (Distances Å)	α	β	γ	δ
Het[AA*] 0000-1111				
Path 1 [OPT without CP]	3.25	3.56	3.43	3.14
Path 2 [OPT then CP]	3.25	3.56	3.43	3.14
Path 3 [OPT with CP]	3.27	3.59	3.46	3.17
Het[BB*] 0001-1110				
Path 1 [OPT without CP]	3.01	3.13	3.10	2.86
Path 2 [OPT then CP]	3.01	3.13	3.10	2.87
Path 3 [OPT with CP]	3.03	3.15	3.12	2.88
Het[CC*] 0010-1101				
Path 1 [OPT without CP]	3.05	3.20	3.18	3.03
Path 2 [OPT then CP]	3.05	3.20	3.18	3.03
Path 3 [OPT with CP]	3.07	3.23	3.21	3.05
Het[DD*] 0011-1100				
Path 1 [OPT without CP]	2.95	3.09	3.08	2.97
Path 2 [OPT then CP]	2.95	3.09	3.08	2.97
Path 3 [OPT with CP]	2.97	3.11	3.10	3.00
Het[EE*] 0100-1011				
Path 1 [OPT without CP]	3.04	3.15	3.18	3.03
Path 2 [OPT then CP]	3.04	3.15	3.18	3.03
Path 3 [OPT with CP]	3.07	3.18	3.20	3.05
Het[FF*] 0101-1010				
Path 1 [OPT without CP]	2.97	3.20	3.20	2.97
Path 2 [OPT then CP]	2.97	3.20	3.20	2.97
Path 3 [OPT with CP]	2.99	3.23	3.23	2.99
Het[GG*] 0110-1001				
Path 1 [OPT without CP]	2.93	3.11	3.11	2.93
Path 2 [OPT then CP]	2.93	3.11	3.11	2.93
Path 3 [OPT with CP]	2.95	3.13	3.13	2.95
Het[HH*] 0111-1000				
Path 1 [OPT without CP]	2.86	3.10	3.12	3.03
Path 2 [OPT then CP]	2.86	3.10	3.12	3.03
Path 3 [OPT with CP]	2.88	3.12	3.14	3.05

A small increase in each of the hydrogen bond lengths across all positions is seen indicating that including CP during geometry optimisation results in the molecular pairs sitting slightly further apart. An increase in bond distance is consistent with that seen in

literature [1]. As a different route is followed to the minimum energy structure in path 3 (opt with CP) the pair geometry differs not only in the final structure but also during the steps on the way to the minimised final structure. To illustrate this the change in energy per optimisation step can be seen for Het[BB*] (Fig. A2.1).

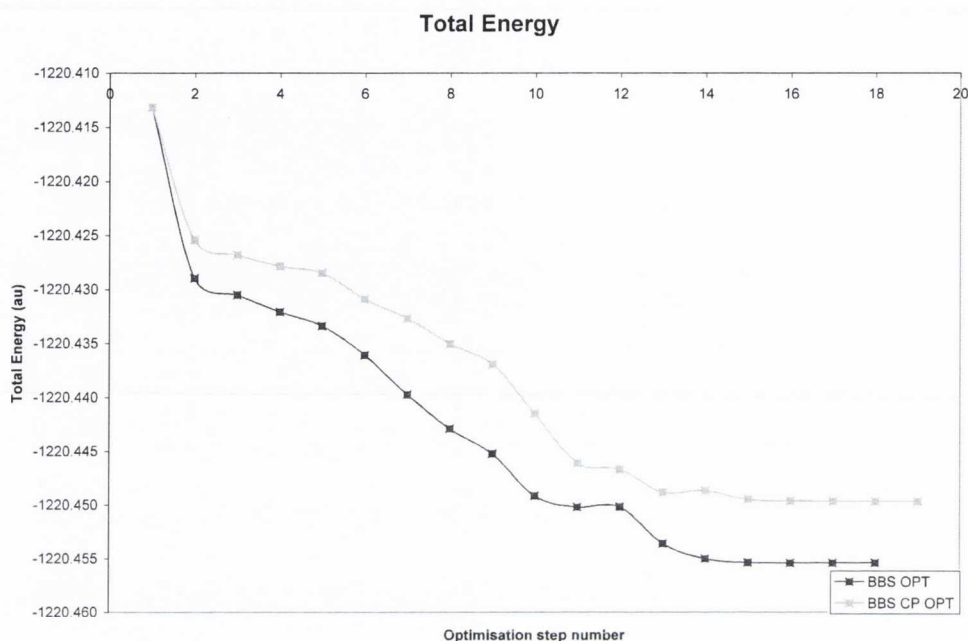


Figure A2.1 Dimer energy and TIE for each of the Het complementary pairs, determined using different bsse correction paths. HF 6-31G(d) basis set.

As expected when BSSE is not removed as in path 1, a curve lower in energy at very optimisation step is observed[2].

1. Van Duijneveldt-van de Rijdt, J. and F.B. Van Duijneveldt, *Convergence to the basis set limit in Ab Initio calculations at the correlated level on the water dimer..* Journal of Chemical Physics, 1992. **97**(7): p. 5019-5030.
2. Salvador, P., *Implementation and application of basis set superposition error-correction schemes to the theoretical modeling of weak intermolecular interactions.* Doctoral thesis 2001: Department of Chemistry and Institute of Computational Chemistry, University of Girona.

A3 Basis set choice: Results

In order to determine how large an effect the choice of basis set has on both energy and molecular structure, a series of test calculations were carried out using Het[DD*] as a test pair (Fig. A3.1).

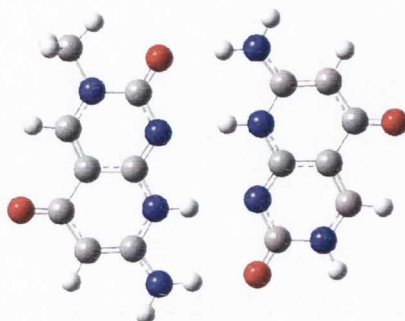


Figure A3.1 Test pair Het[DD*] 0011-1100

The total interaction energy (TIE) (Eqn. AE2.1) for Het[DD*] was calculated using a variety of basis sets and polarisation and diffuse functions. As a larger and larger basis set is used the HF limit (the lowest energy that can be achieved as the basis set nears completeness) should be approached and the changes in structure and energy should become smaller. No geometry constraints were placed on the test pair although the point group was maintained throughout. The results (Table A3.1, Table A3.2) show that in general as the basis set becomes larger the TIE energy for each pair becomes more repulsive (less negative).

$$E_{\text{TIE}}(\text{AB}) = E(\text{AB}) - (E(\text{A}) + E(\text{B})) \quad \text{Equation. A2.1}$$

Table A3.2 TIE calculated for Het[DD*] using different double zeta basis sets.

Double Zeta Valence		
OPT Free Het[DD*]	TIE(kJ/mol)	Number of Basis Function
6-31G	-176.084	271
6-31+G	-169.385	379
6-31G(d)	-147.449	433
6-31G(d,p)	-147.060	475
6-31+G(d)	-138.562	541
6-31+G(d,p)	-137.209	583

Table A3.2 TIE calculated for Het[DD*] using different Triple zeta basis sets

Triple Zeta Valence		
OPT Free Het[DD*]	TIE(kJ/mol)	Number of Basis Function
6-311G	-172.340	393
6-311G(d)	-142.416	528
6-311G(d,p)	-141.731	570
6-311++G(2d,2p)	-127.269	869

This trend of decreasing TIE with increasing basis set makes sense when BSSE is considered (See section 2.3). As the basis set becomes larger the effect of BSSE becomes smaller causing the total interaction energy to decrease. The decreasing trend can be seen below (Fig. A3.2) as a function of the number of basis functions used in each basis set. In general the larger the decrease in energy between different basis sets, the larger the jump in the number of basis functions used.

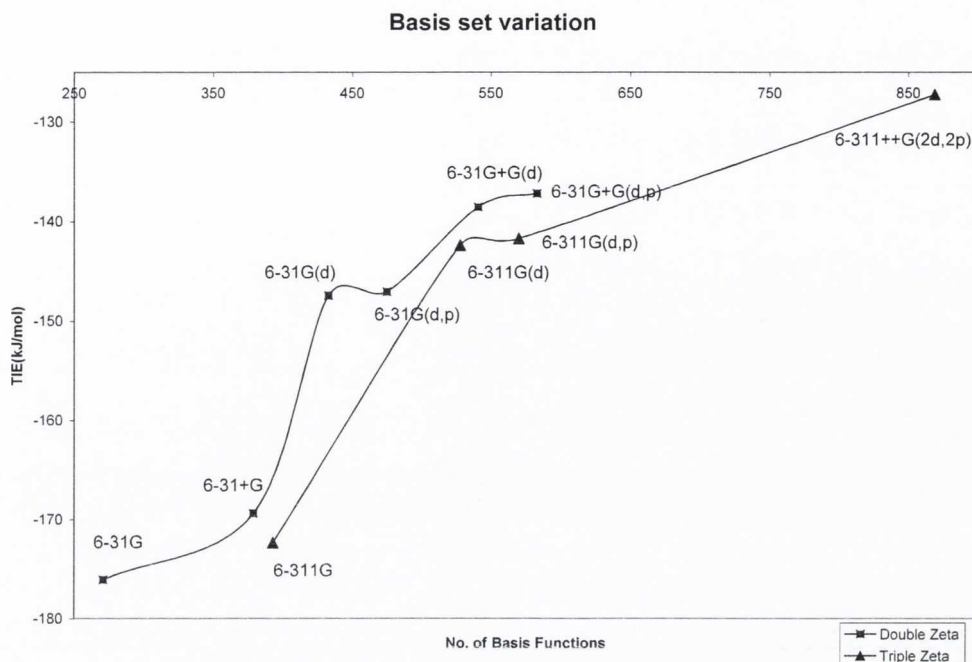


Figure A3.2 TIE calculated for Het[DD*] versus the number of basis functions used.

The overall decrease in TIE on addition of 312 basis functions for double zeta basis sets is 39 kJ/mol and 45 kJ/mol on addition of 476 basis functions for triple zeta basis sets.

Analysis of hydrogen bond distances reveals (Table. A3.3) that as the basis set becomes larger the hydrogen bonds in each position become longer making the two molecules in the pair sit further apart. This trend also agrees with that seen above when the geometry of pairs with and without BSSE were compared.

Figure A3.3 Variation in hydrogen bond distances with basis set for double and triple zeta basis sets

Hydrogen bond distance(Å)				
Double Zeta Valence				
OPT Free Het[DD*]	1	2	3	4
6-31G	2.876	3.016	3.006	2.898
6-31+G	2.889	3.040	3.028	2.912
6-31G(d)	2.954	3.094	3.080	2.972
6-31G(d,p)	2.951	3.090	3.077	2.970
6-31+G(d)	2.960	3.110	3.098	2.984
6-31+G(d,p)	2.960	3.107	3.095	2.984
Triple Zeta Valence				
6-311G	2.873	3.011	3.000	2.895
6-311G(d)	2.960	3.102	3.090	2.984
6-311G(d,p)	2.953	3.093	3.083	2.978
6-311++G(2d,2p)	2.967	3.114	3.102	2.991

The TIE of any pair will change if the basis set used to calculate the energy is changed. A large part of this change will be due to the decrease in BSSE on the use of a more complete basis set. The larger the basis set used the higher (more repulsive) each TIE would become.

Although the use of a different basis sets will change the absolute TIE values, as long as the basis set used is consistent throughout the set of molecules being explored, the overall relative pattern of energies will remain the same.

A4 Average primary and secondary hydrogen bond calculation

An average value for a primary and secondary hydrogen bond can be calculated using a complete set of complementary associations. In doing this all possible combinations of pairs need to be considered. In total when pairs with equal numbers of secondary interactions are removed 15 combinations remain (Table A4.1)

Table A4.1 All combinations of pairs used to determine average an average value for primary and secondary hydrogen bond

0001-1110	0010-0010	0100-1011
0010-0010	0101-1010	0101-1010
0001-1110	0010-0010	0100-1011
0100-1011	0111-1000	0111-1000
0001-1110	0011-1100	0101-1010
0101-1010	0100-1011	0110-1001
0001-1110	0011-1100	0101-1010
0110-1001	0101-1010	0111-1000
0010-0010	0011-1100	0110-1001
0011-1100	0110-1001	0111-1000

The difference between two pairs can be considered as equal to the difference in SI of the pairs (chemical differences or the fact that not all primary and secondary hydrogen bonds are equal are neglected in this model) and from this an average secondary interaction value can be determined. Once a value has been determined it can be used in conjunction with the calculated TIE to find an average primary hydrogen bond value (Fig. A4.2).

Table A4.2 Sample calculation of an average secondary and primary hydrogen bond

Pair	TIE (kJ/mol)	Net SI
0001-1110	-160.883	2
0010-0010	-88.961	-2
Difference	-71.922	4
Average Secondary kJ/mol	-71.992/4	+/-17.981
Average Primary kJ/mol	$17.981*2=35.961+160.888/4$	-31.23

This procedure was repeated for all 15 combinations of pairs and global averages for each bond type found. In all sets of letters AA* 0000-1111 was excluded from the average calculation as it is significantly different from the expected SI patters. In the Zim set, Zim[FF*] 0101-1010 was also excluded due to its large binding interaction energy relative to the other pairs in the set.

A5 The total number of mismatching associations

In order to fully explore the proposed theoretical Het alphabet all possible complementary and non-complementary associations need to be considered. In total 136 possible Het associations exist as detailed in Table A5.1.

Table A5.1 The number of pairs of each category type

Complementary associations	8
Mismatches in one position	32
Mismatches in two positions	48
Mismatches in three positions	32
Mismatches in four positions	16

The XNOR (Table A5.2) function is a usefully tool in determining the number of mismatches of each type that exist.

Table A5.2 XNOR truth table

Bit 1	1	1	0	0
Bit 2	0	1	0	1
XNOR	0	1	1	0

The complementary Hamming distance \bar{d} of an XNOR product can be used to determine in how many positions a pair of molecules mismatch. In total 16 XNOR values exist and these can be broken into categories based on the \bar{d} value of each (Table A5.3).

Table A5.3 All possible XNOR values and the weight (no. of mismatches) present in each

XNOR	$\bar{\partial}$
0000	0
0001	1
0010	1
0100	1
1000	1
0011	2
0101	2
0110	2
1100	2
0101	2
1010	2
0111	3
1110	3
1011	3
1101	3
1111	4

A5.1 Mismatches in two positions

For pairs with two mismatches, two out of four positions will match giving an XNOR with two 0s (one in each position that matches) and two 1s (in the mismatching positions). For example;

	0011
	<u>0110</u>
XNOR	1010

In order to calculate the number of mismatches in two position three factors must be taken into account;

1. The number of arrangements of positions in which the desired number of mismatch can occur
2. The number of arrangements of types of mismatches that can occur
3. The number of arrangements of matching positions that can exist

The calculation method used to determine the total number of pairs that can mismatch in two positions is shown here to explain each of the three components necessary. The total number of pairs with fewer or greater mismatches than two can also be determined using this methodology. The first step is to determine how many arrangements of the two mismatching positions exist. This can be calculated using the mathematical formula for combinations;

$$c(n,r) = \frac{n!}{r!(n-r)!} \quad \text{Equation A5.1}$$

In this formula n is the total number of positions (4) and r the number of these positions we wish to choose from. r will be equal to the number of mismatches (in the case of two mismatches this will be two). In total there are four possible hydrogen bonding positions between associations and we need to consider all possibilities in which two of these will mismatch. Filling in the relevant numbers to the equations gives;

$$c(n,r) = \frac{4!}{2!(4-2)!} = \frac{4!}{2!(2!)} = \frac{24}{4} = 6$$

This means that if any two out of a four positions mismatch, 6 possible choices of which two positions mismatch exist. Using the same labeling system shown in (Fig. A5.1), the six unique combinations would be $\alpha\beta$, $\gamma\delta$, $\beta\gamma$, $\alpha\delta$, $\alpha\gamma$, $\beta\delta$

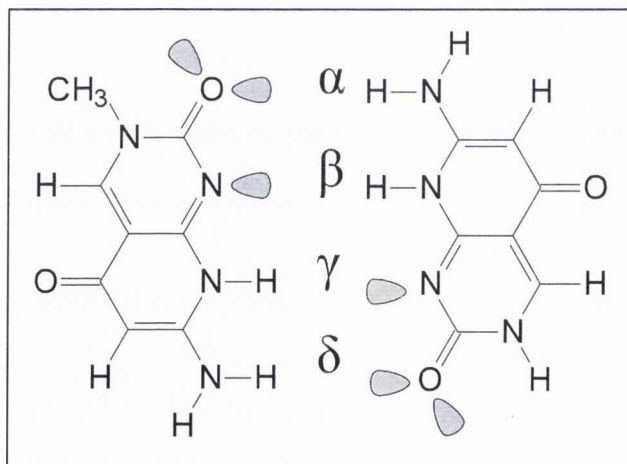


Figure A5.1 Het[DD*] showing the four hydrogen bonding positions labelled α β γ δ .

The second step is to determine how many different groupings of mismatch type can exist. Each mismatch could be either H-H or Lp-Lp in type. When two mismatches are present they could either be all of one type or one of each. The total number of possible arrangements can be calculated using $2^b=N$, as only two types of mismatch exist. b is the number of mismatches to be picked and N is the total number of possibilities. Using the formula and applying it to associations that mismatch in two positions gives $N=4$. The full details of all possibilities for this can be seen in the table shown below (Table A5.3).

Table A5.3 All four possible arrangements of mismatches

1	H-H H-H
2	Lp-Lp Lp-Lp
3	H-H Lp-Lp
4	Lp-Lp H-H

The final step is to determine how many (polarity) arrangements of matching positions remain. For every pair with a single mismatch three positions will match. The total number of arrangements can be determined using $\text{Log}_2N=2$ and solving for N , which in this case is 4. Although 4 possible arrangements of binary digits exist they do not occur independently but in pairs (as we are considering the interaction between pairs of letters), giving a total number of arrangements of 2. The total number of mismatches in one position can now be arrived at by multiplying the three necessary components together;

$$6(\text{positions}) * 4(\text{types}) * 2(\text{complementary arrangements}) = 48$$

A5.2 Mismatches in one position

The number of positions in which a single mismatch can occur is given by taking the overall number of hydrogen bonding positions 4 and determining how many combinations of 1 can be made from the four positions (although with one mismatch this is a trivial example the same procedure is followed when higher number of mismatches are present) using the standard mathematical formula for combinations (Eqn. A5.1). In this equation n refers to the total number of hydrogen bonding positions present (4) and r to the number in

which we have a mismatch (in this case 1). In total there are four positions and we need to chose one. Filling in the relevant numbers to the formula (Eqn A5.1) gives;

$$c(n,r) = \frac{4!}{1!(4-1)!} = \frac{24}{6} = 4$$

There are four positions in which a single mismatch can occur. The next step is to calculate the number of arrangements of mismatch types that can be present. This can be done by raising the number of mismatch types (always 2 as only H-H or Lp-Lp are mismatches) to the power of the number of mismatching positions. In the case of one mismatch doing so gives $2^1=2$. The final step is to determine how many arrangements of matching positions remain. For every pair with a single mismatch three positions will match. The total number of arrangements can be determined using $\text{Log}_2N=3$ and solving for N, which in this case is 8. Although 8 possible arrangements of binary digits exist they do not occur independently but in pairs (as we are considering the interaction between pairs of letters), giving a total number of arrangements of 4. The total number of mismatches in one position can now be arrived at by multiplying the three necessary components together;

$$4(\text{positions}) * 2(\text{types}) * 4(\text{complementary arrangements}) = 32$$

The table below shows how total number of pairs that can be formed for a given mismatch broken down based on mismatch type and each of the three calculation components needed (Table A5.4).

Table A5.4 Total number of mismatches for each mismatch type. Data broken down into number of mismatches and calculation components needed

	One Mismatch	Two Mismatches	Three Mismatches
The number of arrangements of positions in which the desired number of mismatches can occur	4	6	4
The number of arrangements of types of types of mismatches that can occur	2	4	8
The number of arrangements of matching positions that can exist	4	2	1
Total Number of Pairs	32	48	32

A6 Standard deviation

In order to gain an insight in the variance of a series of data from its mean value the standard deviation (population variance) can be determined using Eqn. A6.1. In this equation x_i represents each individual data point in the sample, μ the average of all members in the sample and N the total number of data points in the sample [1].

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Equation A6.1

1. Mendenhall W, Beaver B, *Introduction to probability and statistics*. Thirteenth Edition ed. 1999: Brooks/ColePg. 62

A7 Calculation time

The table below lists some average calculation times for the associations studied in this thesis. It is noted that due to job queuing time ranging from minutes to days depending on cluster demand the overall time taken for a calculation is longer than that simply listed here. The time quoted here do not take into account calculations that failed to converge during the optimization cycle and so needed to be run again.

Table A7.1 Sample calculation running times. See A8 for Full opt explanation

Pair	Method	CPU time	No. of processors
Het[DD*]	HF 6-31G(d)	5 Hrs 17 Mins	2
Het[HH*]	HF 6-31G(d)	11 Hrs 53 Mins	2
Het[B*H*]	HF 6-31G(d)	2 Hrs 8 Mins	2
Zim[DA*]	HF 6-31G(d)	17 hrs 5 Mins	4
Het[CC*]	MP2 6-31G(d)	25 Hrs 5 Mins	8
Het[DD*]	MP2 6-31G(d)	31 Hrs 12 Mins	8
	MP2 6-31G(d) Full		
Het[CC*]	Opt	88 Hrs 16 Mins	8
Het[A*A*]	SP MP2 6-31G(d)	4 Hrs 3 Mins	4

A8 MP2-Further considerations

The data shown in this section is for MP2 with a frozen core (this is the default setting in Gaussian 03W). A frozen core means that correlation is only included for the outer-shells during calculation. In order to assess how big an effect this could have on results some test calculations were performed using full MP2 (Table A8.1)

Table A8.1 MP2 Full data for test pairs

	TIE OPT MP2 (kJ/mol)	TIE OPT MP2 Full (kJ/mol)	Difference (kJ/mol)
Het[AA*]	-97.769	-99.369	1.599
Het[CC*]	-131.154	-133.636	2.481
Het[EE*]	-137.337	-139.916	2.579
Het[GG*]	-173.818	-176.689	2.870
Het[DH*]	-133.232	-135.087	1.855
Het[B*D]	-71.063	-71.556	0.493
Het[BD*]	-134.626	-136.490	1.864
Het[E*F]	-39.948	-42.376	2.428
Het[AE]	-24.350	-25.125	0.774
Het[CF*]	58.489	56.463	2.026
Het[C*F]	91.491	88.634	2.857
Het[A*C*]	166.568	163.423	3.145

Only a small difference is seen in terms of the relative interaction energies. If MP2 optimizations had been performed including correlation on the core shells a difference would have been seen in the absolute energies calculated and in the averages determined but this would not be enough to cause a change in the overall results.

A9 Zimmerman pair set redundancies

In the full Zim set of molecules Zim[BB*] and Zim[HH*] are both represented by the same molecular pair (Fig. A9.1), one pattern can be converted into the other flipping the pair vertically. This duplication is possible in this case because no anchor or backbone structure is (in this initial investigation) in place. If an anchoring group was attached to each pair at a specific position Zim[HH*] and Zim[BB*] would no longer be identical in molecular representation and using identical molecules to represent more than one pattern would no longer be possible. The same applies to pairs Zim[CC*] and Zim[EE*]. Three of the molecules in the Zim set are symmetrical in shape and pattern, ZimA* 1111, ZimG 0110 and ZimG* 1001, meaning that simplifications in the number of calculations that need to be performed for a complete set of results can be made.

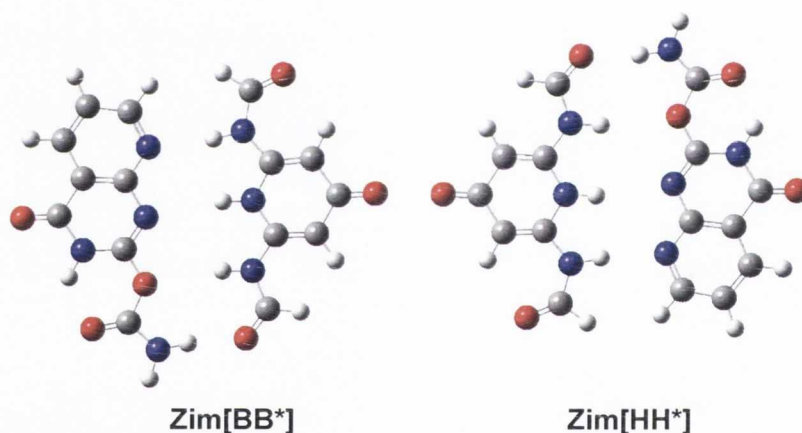


Figure A9.1 Zim[BB*] Zim[HH*], represented by the same molecules but in a different orientation.

Although the full set of results will always be shown for Zim pairs to allow complete analysis and direct comparison to the Het set, pattern simplifications will be used wherever possible in the Zim molecular set.

It is not always possible to apply STRD conditions to mismatching associations in the Zimmerman alphabet, Zim[AH*] (Fig. A9.2) for example has a Lp-Lp mismatch in beta position. Both ZimA and ZimH* have a flexible chain-like structure in the beta position thus making it impossible to lock an angle straight across of 180°. In this specific case the angle was left free. This is a rare situation and that only arises in this specific instance

and in Zim[AA], Zim[BB] and Zim[H*H*] which mismatch in four positions, it does however highlight the increased flexibility that can exist due to a mixture of chain and rigid parts being used in the Zimmerman molecules.

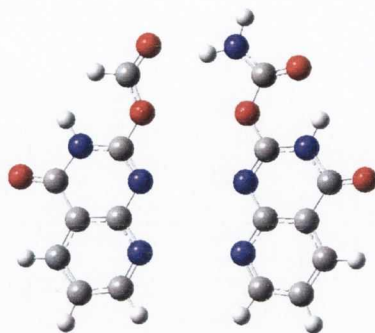


Figure A9.2 Zim[AH*] 0000-1000

Bibliography

- Arunan, E., et al., *Definition of a Hydrogen Bond*. IUPAC task group, 2010.
- Beijer, F.H., et al., *Self-complementarity achieved through quadruple hydrogen bonding*. *Angewandte Chemie-International Edition*, 1998. **37**(1-2): p. 75-78.
- Beijer, F.H., et al., *Strong dimerization of ureidopyrimidones via quadruple hydrogen bonding*. *Journal of the American Chemical Society*, 1998. **120**(27): p. 6761-6769.
- Biggs, N., *Discrete Mathematics (second edition)*. OUP, 2002.
- Blight, B.A., et al., *An AAAA-DDDD quadruple hydrogen-bond array*, *Nature Chemistry*, 2011. **3**(3): p. 244-248.
- Boys, S.F. and F. Bernardi, Calculation of small molecular interactions by differences of separate total energies - some procedures with reduced errors. *Molecular Physics*, 1970. **19**(4): p. 553-&.
- Dawkins, R., *The Blind Watchmaker*. Originally published by Longman Scientific & Technical 1986, 1991: p. 112.
- Desiraju, G.R., Steiner, T, *The Weak Hydrogen Bond, In Structural Chemistry and Biology*. 1999: OUP.
- Dewar, M.J.S., et al., *AMI - A new general purpose quantum mechanical molecular model*. *Journal of the American Chemical Society*, 1985. **107**(13): p. 3902-3909.
- Dickerson, R.E., et al., *The anatomy of A-DNA, B-DNA, and Z-DNA*.
- Dong, F. and R.E. Miller, *Vibrational transition moment angles in isolated biomolecules: A structural tool*. *Science*, 2002. **298**(5596): p. 1227-1230.
- Einstein, A., in *Festschrift fur Aurel Stodola*, . E. Honegger, Ed. Vol. Orell Fussli Verlag, Zurich. 1929
- Eschenmoser, A. and M. Dobler, *Why pentose and not hexose nucleic-acids. 1. Introduction to the problem, conformational-analysis of oligonucleotide single strands containing 2',3'-dideoxylucopyranosyl building-blocks(homo-DNA), and reflections on the conformation of A-DNA and B-DNA*.
- Eschenmoser, A., *Chemical etiology of nucleic acid structure*. *Science*, 1999. **284**(5423): p. 2118-2124.
- Eschenmoser, A., *Hexose Nucleic-Acid*. *Pure and Applied Chemistry*, 1993. **65**(6): p. 1179-1188.
- Evans, T.A. and K.R. Seddon, *Hydrogen bonding in DNA - a return to the status quo*. *Chemical Communications*, 1997(21): p. 2023-2024.
- Fock, V., *Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems*. *Zeitschrift fuer Physik*, 1930. **62**.
- Frisch, M.J., et al., *Extensive theoretical-studies of the hydrogen-bonded complexes (H2O)2, (H2O)2H+, (HF)2, (HF)2H+, F2H-, AND (NH3)2*. *Journal of Chemical Physics*, 1986. **84**(4): p. 2279-2289.
- Galano, A. and J.R. Alvarez-Idaboy, *A new approach to counterpoise correction to BSSE*. *Journal of Computational Chemistry*, 2006. **27**(11): p. 1203-1210.
- Grabowski, S.J., *Ab initio calculations on conventional and unconventional hydrogen bonds - Study of the hydrogen bond strength*. *Journal of Physical Chemistry A*, 2001. **105**(47): p. 10739-10746.
- Guckian, K.M., T.R. Krugh, and E.T. Kool, *Solution structure of a nonpolar, non-hydrogen-bonded base pair surrogate in DNA*. *Journal of the American Chemical Society*, 2000. **122**(29): p. 6841-6847.
- Guo, H. and M. Karplus, *Ab Initio studies of polyenes.1. 1,3-butadiene*. *Journal of Chemical Physics*, 1991. **94**(5): p. 3679-3699.

- Hamming, R.W., *Error detecting and error correcting codes*. Bell System Technical Journal, 1950. **29**(2): p. 147-160.
- Hartree, D., *The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods*. Proceedings of the Cambridge Philosophical Society, 1928. **28**.
Helvetica Chimica Acta, 1992. **75**(1): p. 218-259.
- Herbert, H.E., et al., *Hydrogen-bonding interactions in peptide nucleic acid and deoxyribonucleic acid: A comparative study*. Journal of Physical Chemistry B, 2006. **110**(7): p. 3336-3343.
- Hobza, P. and J. Sponer, *Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations*. Chemical Reviews, 1999. **99**(11): p. 3247-3276.
- Humphreys, J.F., Prest, M.Y., *Numbers, Groups & Codes (second edition)*. CUP, 2004.
into DNA and RNA. Journal of the American Chemical Society, 1989. **111**(21): p. 8322-8323.
- Jeffrey, G.A., *An Introduction to Hydrogen Bonding*. 1997: OUP.
- Jensen, F., *Introduction to Computational chemistry*. 2007: Wiley.
- Johansson, A., P. Kollman, and Rothenbe, S., *Application of functional Boys Bernardi counterpoise method to molecular potential surfaces*. Theoretica Chimica Acta, 1973. **29**(2): p. 167-172.
- Jorgensen, W.L. and J. Pranata, *Importance of Secondary Interactions in Triply Hydrogen-Bonded Complexes - Guanine-Cytosine Vs Uracil-2,6-Diaminopyridine*. Journal of the American Chemical Society, 1990. **112**(5): p. 2008-2010.
- Kool, E.T., J.C. Morales, and K.M. Guckian, *Mimicking the structure and function of DNA: Insights into DNA stability and replication*. Angewandte Chemie-International Edition, 2000. **39**(6): p. 990-1009.
- Kubicki, J.D., G.A. Blake, and S.E. Apitz, *Molecular orbital calculations for modeling acetate-aluminosilicate adsorption and dissolution reactions*. Geochimica Et Cosmochimica Acta, 1997. **61**(5): p. 1031-1046.
- Kurita, N., V.I. Danilov, and V.M. Anisimov, *The structure of Watson-Crick DNA base pairs obtained by MP2 optimization*. Chemical Physics Letters, 2005. **404**(1-3): p. 164-170.
- Leach, A.R. and P.A. Kollman, *Theoretical investigations of novel nucleic-acid bases*. Journal of the American Chemical Society, 1992. **114**(10): p. 3675-3683.
- Leach, A.R., *Molecular Modelling Principles and Applications (second edition)*. Prentice Hall, 2001
- Lehn, J., Marie, *Supramolecular Chemistry, Concepts and Perspectives* 1995: VCH
- Leszczynski, J. *Are the amino groups in the nucleic acid bases coplanar with the molecular rings Ab Initio HF 6-31G* and MP2 6-31G* studies*. 32nd Sanibel International Symp on the Application of Fundamental Theory to Problems of Biology and Pharmacology. 1992. St Augustine, Fl: John Wiley & Sons Inc.
- Li, X.Q. and P. Fan, *A duplex DNA model with regular inter-base-pair hydrogen bonds*. Journal of Theoretical Biology, 2010. **266**(3): p. 374-379.
- Liang, W., et al., *Systematic theoretical investigations on all of the tautomers of guanine: From both dynamics and thermodynamics viewpoint*. Chemical Physics, 2006. **328**(1-3): p. 93-102.
- Luisi, B., et al., *On the potential role of the amino nitrogen atom as a hydrogen bond acceptor in macromolecules*. Journal of Molecular Biology, 1998. **279**(5): p. 1123-1136.

- Lukin, O. and J. Leszczynski, *Rationalizing the strength of hydrogen-bonded complexes. Ab initio HF and DFT studies.* Journal of Physical Chemistry A, 2002. **106**(29): p. 6775-6782
- Mac Donaill, D.A., *A parity code interpretation of nucleotide alphabet composition.* Chemical Communications, 2002(18): p. 2062-2063.
- Mac Donaill, D.A., *Molecular Error-Coding: Why Nucleotides Come in Two Sizes In preparation.*
- Mac Donaill, D.A., *Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition.* Origins of Life and Evolution of the Biosphere, 2003. **33**(4-5): p. 433-455.
- Mac Donaill, D.A. and D. Brocklebank, *An ab initio quantum chemical investigation of the error-coding model of nucleotide alphabet composition.* Molecular Physics, 2003. **101**(17): p. 2755-2762.
- Mendenhall W, B.R., Beaver B, *Introduction to probability and statistics.* Thirteenth Edition ed. 1999: Brooks/ColePg. 62
- Miller, S.L., *A Production of amino acids under possible primitive earth conditions.* Science, 1953. **117**(3046): p. 528-529.
- Mirzaei, M. and N.L. Hadipour, *A computational NQR study on the hydrogen-bonded lattice of cytosine-5-acetic acid.* Journal of Computational Chemistry, 2008. **29**(5): p. 832-838.
- Møller, C. and M.S. Plesset, *Note on an Approximation Treatment for Many-Electron Systems.* Physical Review, 1934. **46**(7): p. 618.
- Murray, T.J. and S.C. Zimmerman, *New Triply Hydrogen-Bonded Complexes with Highly Variable Stabilities.* Journal of the American Chemical Society, 1992. **114**(10): p. 4010-4011.
- Neidle, S., *Nucleic acid structure and recognition.* OUP, 2002.
- Piccirilli, J.A., et al., *Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet.* Nature, 1990. **343**(6253): p. 33-37
- Popelier, P.L.A. and L. Joubert, *The elusive atomic rationale for DNA base pair stability.* Journal of the American Chemical Society, 2002. **124**(29): p. 8725-8729.
- Pranata, J., S.G. Wierschke, and W.L. Jorgensen, *Opls Potential Functions for Nucleotide Bases - Relative Association Constants of Hydrogen-Bonded Base-Pairs in Chloroform.* Journal of the American Chemical Society, 1991. **113**(8): p. 2810-2819.
- Quinn, J.R. and S.C. Zimmerman, *With regard to the hydrogen bonding in complexes of pyridylureas, less is more. A role for shape complementarity and CH...O interactions?* Organic Letters, 2004. **6**(10): p. 1649-1652.
- Quinn, J.R., et al., *Does the A.T or G. C base-pair possess enhanced stability? Quantifying the effects of CH...O interactions and secondary interactions on base-pair stability using a phenomenological analysis and ab initio calculations.* Journal of the American Chemical Society, 2007. **129**(4): p. 934-941.
- Riggs, N.V., *An Ab Initio study of the stationary structures of the major gas phase tautomer of adenine.* Chemical Physics Letters, 1991. **177**(4-5): p. 447-450.
- Roberts, C., R. Bandaru, and C. Switzer, *Theoretical and experimental study of isoguanine and isocytosine: Base pairing in an expanded genetic system.* Journal of the American Chemical Society, 1997. **119**(20): p. 4640-4649.
- Sadlej, A.J., *Exact perturbation treatment of the basis set superposition correction.* Journal of Chemical Physics, 1991. **95**(9): p. 6705-6711.

- Salvador, P., D. Asturiol, and I. Mayer, *A general efficient implementation of the BSSE-free SCF and MP2 methods based on the Chemical Hamiltonian Approach*. Journal of Computational Chemistry, 2006. **27**(13): p. 1505-1516.
- Salvador, P., *Implementation and application of basis set superposition error-correction schemes to the theoretical modeling of weak intermolecular interactions*. Doctoral thesis 2001: Department of Chemistry and Institute of Computational Chemistry, University of Girona.
- Salvador, P., *Implementation and application of basis set superposition error-correction schemes to the theoretical modeling of weak intermolecular interactions*. Doctoral thesis 2001: Department of Chemistry and Institute of Computational Chemistry, University of Girona.
- Sartorius, J. and H.J. Schneider, *A general scheme based on empirical increments for the prediction of hydrogen-bond associations of nucleobases and of synthetic host-guest complexes*. Chemistry-a European Journal, 1996. **2**(11): p. 1446-1452.
- Schwenke, D.W. and D.G. Truhlar, *Systematic study of basis set superposition errors in the calculated interaction energy of 2 HF molecules*. Journal of Chemical Physics, 1985. **82**(5): p. 2418-2426.
- Sherrington, D.C. and K.A. Taskinen, *Self-assembly in synthetic macromolecular systems via multiple hydrogen bonding interactions*. Chemical Society Reviews, 2001. **30**(2): p. 83-93.
- Sijbesma, R.P. and E.W. Meijer, *Quadruple hydrogen bonded systems*. Chemical Communications, 2003(1): p. 5-16.
- Simon, S., M. Duran, and J.J. Dannenberg, *How does basis set superposition error change the potential surfaces for hydrogen bonded dimers?* Journal of Chemical Physics, 1996. **105**(24): p. 11024-11031.
- Sinden, R.R., *DNA structure and function*. 1994: Academic Press
- Sponer, J. and P. Hobza, *Bifurcated hydrogen bonds in DNA crystal structures- An ab initio quantum chemical study*. Journal of the American Chemical Society, 1994. **116**(2): p. 709-714.
- Sponer, J. and P. Hobza, *DNA base amino groups and their role in molecular interactions: Ab initio and preliminary density functional theory calculations*. International Journal of Quantum Chemistry, 1996. **57**(5): p. 959-970.
- Sponer, J. and P. Hobza, *Nonplanar geometry of DNA bases- Ab Initio 2nd order Moller-Plesset study*. Journal of Physical Chemistry, 1994. **98**(12): p. 3161-3164.
- Stewart, J.J.P., *Optimization of parameters for semiempirical methods.1. Method*. Journal of Computational Chemistry, 1989. **10**(2): p. 209-220.
- Switzer, C., S.E. Moroney, and S.A. Benner, *Enzymatic incorporation of a new base pair* Szathmary, E., *4 Letters in the genetic alphabet- A frozen evolutionary optimum*. Proceedings of the Royal Society of London Series B-Biological Sciences, 1991. **245**(1313): p. 91-99.
- Szathmary, E., *What is the optimum size for the genetic alphabet*. Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(7): p. 2614-2618.
- Van Duijneveldt-van de Rijdt, J. and F.B. Van Duijneveldt, *Convergence to the basis set limit in Ab Initio calculations at the correlated level on the water dimer..* Journal of Chemical Physics, 1992. **97**(7): p. 5019-5030.
- Wachtershauser, G., *Evolution of the first metabolic cycles*. Proceedings of the National Academy of Sciences of the United States of America, 1990. **87** : P. 200-204.
- Wang, S.Y. and H.F. Schaefer, *The small planarization barriers for the amino group in the*

- nucleic acid bases*. Journal of Chemical Physics, 2006. **124**(4): p. 8.
- Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids- A structure for deoxyribose nucleic acid*. Nature, 1953. **171**(4356): p. 737-738.
- Wilson, A.J., *Hydrogen bonding Attractive arrays*. Nature Chemistry, 2011. **3**(3): p. 193-194.
- Wilson, A.J., *Non-covalent polymer assembly using arrays of hydrogen-bonds*. Soft Matter, 2007. **3**(4): p. 409-425.
- Yockey, H.P., *Information Theory and Molecular Biology*. CUP, 1992: p. 102.
- Zheng, X.Y., et al., *Density Functional Theory Study of the Free and Tetraprotonated Spheroidal Macrotricyclic Ligands and the Complexes with Halide Anions: F-, Cl-, Br*. Journal of Computational Chemistry. **31**(4): p. 871-881.
- Zimmerman, S.C. and F.S. Corbin, *Heteroaromatic modules for self-assembly using multiple hydrogen bonds*. Molecular Self-Assembly, 2000. **96**: p. 63-94.