High-speed Molecular Mechanics Searches for Optimal DNA Interaction Sites

Ingrid Lafontaine and Richard Lavery*

Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13, rue Pierre et Marie Curie, Paris 75005, France

Abstract: We have recently developed a theoretical means of studying the mechanical and interaction properties of nucleic acids as a function of their base sequence. This approach, termed ADAPT, can be used to obtain the physical properties of millions of base sequences with only modest computational expense. ADAPT is based on a multi-copy algorithm using special nucleotides ("lexides") containing all four standard bases whose contribution to the energy of the molecule can be varied. We present here a deeper study of the energy minima which occur in the multi-dimensional space defined by these variable sequences. We also present an extension of the approach termed "gene threading" which enables us to scan genomic sequence data in an attempt to locate preferential binding sites. This technique is illustrated for the case of TATA-box protein binding. ADAPT enables us to demonstrate that, for this protein, DNA deformation alone explains a large part of the experimentally observed consensus binding sequence.

Keywords: protein-DNA recognition, deformation energy, binding sites, multi-copy algorithm, genome analysis

INTRODUCTION

Over the last few years whole genome sequencing has provided biologists with an unprecedented amount of data on an increasingly wide variety of living organisms. However, much of the information contained in these data still remains to be extracted and developing the tools necessary for this extraction represents one of the major challenges facing both bioinformatics and theoretical biochemistry today. While certain features of genome sequences, such as coding regions, are relatively easy to identify, the control regions linked to transcription factors binding are much more difficult to find [1], and we have chosen to focus on this problem.

The fact that many such factors have poorly defined consensus sequences suggests that if we are to identify their binding sites it will be necessary to go beyond simple sequence motif searches and to calculate the mechanical or even dynamic properties of the targeted base sequences [2]. Such data would also be useful for understanding how DNA is packaged within the cell and how it responds to mechanical or topological stress. Although detailed molecular simulations have reached the stage where such properties can be predicted for short fragments of DNA [3], we are faced with the difficulty of obtaining the necessary data sufficiently rapidly to be able to predict the properties of sequences containing many millions of base pairs in reasonable computation times. This difficulty has led to most attempts in this direction being limited to empirical approaches where predictions are based on tables of, hopefully, additive, dinucleotide or trinucleotide properties [2, 4].

We are attempting to develop an alternative solution to this problem which retains an all-atom model of DNA and does not assume that sequences will be built up from overlapping fragments. In a recent publication [5] we have described how base sequence can be made variable within an all-atom model by constructing the DNA with special nucleotides which contain all four standard bases. By varying coefficients which control the presence of each such base, it is possible to optimize sequences in the same way that molecular mechanics is generally used to optimize molecular conformations. We have already used this methodology to show that we can rapidly find sequences which favor given DNA conformations or given DNA interactions. The specific examples we have already treated involved the B-Z transition, intrinsically curved DNA sequences and simple ligand binding [5].

We now turn to the genetically more important problem of protein binding. To demonstrate the possibilities of our new approach we have chosen the TATA-box binding protein (TBP) which, as part of the transcription factor TFIID, plays a crucial role in forming the transcription initiation complex in eukaryotic organisms [6]. We will limit ourselves here to studying the sequence dependence of the structural deformation induced in DNA when TBP binds. However, this is known to be an important aspect of the TBP complexation, both because the deformation is large and because the protein forms only few direct recognition contacts between its amino acid side chains and the bases constituting its target site. Indeed both experimental [7-9] and theoretical studies [10-13] have already shown that TBP affinity for a given site can be related both to the intrinsic conformation and to the dynamic properties of the base sequence forming the site. Specifically, TBP binds on the minor groove face of DNA, dramatically opening this groove, compressing the opposing major

*Address correspondence to this author at the Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13, rue Pierre et Marie Curie, Paris 75005, France; E-mail: rlavery@ibpc.fr

1386-2073/01 $28.00+.00

groove and bending the DNA by almost 100° away from the protein [14-17]. These changes are favored either by sequences which are already bent in the correct direction or which are sufficiently flexible to deform in the desired way.

We demonstrate in this article how our approach can be used to locate such sequences by energy minimization or by combinatorial sequence searches. We also show that we can search for likely binding sites within genomic sequences using a technique related to so-called "threading". In the protein field, “threading” is used to identify the 3D folds which are likely to be adopted by unidentified amino acid sequences [18]. We also make a more detailed study of the energy hypersurface in what we can term "sequence space". The results obtained show that TBP binding sites are indeed strongly related to DNA deformation properties and also suggest that this new technique may become an important tool for locating a range of protein binding sites.

METHODOLOGY

The starting point of our methodology is the creation of special nucleotides, termed lexides, which contain all four bases thymine (or uracil for RNA), cytosine, adenine and guanine (T or U, C, A, G) linked to a single sugar C1' atom [5]. Since our calculations are carried out in an internal coordinate framework where nucleic acid flexibility is limited to single bond torsion angles and sugar and backbone valence angles [19, 20], the bases are rigid bodies (with the exception of the thymine C5 methyl torsion) and they remain superposed throughout the procedure.

Lexides are incorporated into energy calculations by attributing a variable coefficient C_{ik} to each base k belonging to lexide i. These coefficients, which are normalized for each lexide ($\Sigma_{k=1,4}$ C_{ik} = 1), determine the contribution of each base to the total energy of the molecule by simply multiplying each of the normal energy terms in the force field. Therefore, if a given pairwise interaction involves a lexide base atom and a backbone atom (or any atom belonging to a normal nucleotide), two lexide base atoms, the energy term will be multiplied by two different coefficients. It should be noted that bases belonging to the same lexide do not interact. It is also possible to simplify the treatment of base pairs by only allowing paired bases within the corresponding pair of lexides to interact and by using a single set of coefficients (each base being coupled with its complementary base) for this lexide pair. Note that the latter approach is only slightly different from that described in our earlier publication [5], but it allows faster calculations and avoids the need for a special energy normalization constant in the case of paired lexides.

Lexides have been added to the nucleotide library used by the JUMNA program which can energy minimize the conformation of nucleic acids using a combination of helical and internal coordinates [19, 20]. When they are included in a DNA or RNA molecule, it becomes possible to study a new range of sequence effects. If, for a given lexide, we set a single coefficient to 1.0, the others becoming 0.0 by the normalization condition, we have simply created a standard nucleotide. If, on the other hand, we set the coefficients of A and G to 0.5, we can calculate with a canonical purine nucleotide (R). Alternatively, setting all the lexide coefficients to 0.25 creates an average nucleotide (N) where all sequence effects are averaged out.

One may imagine energy minimizing a base sequence simply using the lexide coefficients as variables. However, this would imply changing the chemical composition of the molecule being studied and would make energy minimization impossible since molecular mechanics force fields calculate conformational rather than formation energies. It is however possible to overcome this objection by comparing two different states of the same molecule, for example, two different conformations. In this case energy minimization in sequence space becomes possible and should locate the sequence which will minimize the energy difference between the two conformations in question.

JUMNA is used to create the two conformations we wish to compare. In the case of the present study our aim was to locate sequences favoring TBP binding. We therefore used a target conformation deformed to create a TBP binding site and a canonical B-DNA reference conformation. Both conformations were built with lexides and were energy minimized using an "average" base sequence, where all lexide coefficients have been set to 0.25. This guarantees that the resulting conformations are sterically compatible with all possible base sequences. To create the TBP binding site we used the CONTACT program [21] which restrains the atoms of DNA belonging to the experimental protein-DNAinterface to adopt the same relative positions in the model DNA as they do in the experimental complex. In this case, we used the 1.9 Å resolution crystallographic conformation of the Human TATA-box binding protein with a 16 nucleotide long DNA fragment [17, PDB reference 1CDW]. DNA atoms belonging to the protein-DNA interface were defined using a 3.5 Å cutoff between pairs of DNA and protein atoms, and the relative positions of the interface atoms were permitted to move within spheres of 0.4 Å radius to allow for the limited flexibility of the JUMNA internal coordinate model and for small imprecisions in the experimental data. This procedure creates what can be termed a "molecular mold" which can then be applied to any fragment of DNA. We have used this flexibility to create a TBP binding site within three DNA fragments containing respectively 10, 16 and 24 base pairs. In each case, the binding site is centered within the fragment, occupying respectively base pairs 2-8, 5-11 and 9-15. The non-hydrogen atoms forming these base pairs have RMS fits to the crystallographic data of roughly 1 Å, and the subset of interface atoms have RMS fits of roughly 0.4 Å.

Following energy minimization, we create energy matrices describing the optimal conformations. These matrices group together all energy terms which are multiplied by two, one or no lexide coefficients. Thus, all the phosphodiester backbone atom interactions are stored in a single element which is independent of the lexide coefficients. Base (lexide)-backbone interactions become the diagonal of the matrix, whose dimension is equal to the number of lexides in the DNA fragment. Finally, base-base interactions form the upper triangle of the matrix. Note that since each lexide contains four bases, the matrix is 16 elements in depth for off-diagonal elements, these elements containing the energies of individual base-base interactions : TT, TC, TA, TG, AT, ..., GG. The diagonal elements of the matrix are 4 elements in depth (one for each of the lexide bases, T, C, A and G).

These matrices are used by a new program, ADAPT [5], to finally carry out energy minimization in sequence space. The variables for this procedure are the lexide coefficients. They specify the base sequence which will be common to both the target and reference DNA conformations. Note that this sequence may be made up of "pure" (that is, with a single lexide coefficient equal to 1.0, the others being 0.0) or "mixed" bases depending on the lexide coefficients. For any given set of coefficients the energy of each DNA conformation can be calculated by simply multiplying the matrix elements by the appropriate coefficient and summing the results. Similarly, the energy difference between the two conformations, which constitutes the target function for the minimizer, can be obtained quickly by subtracting the two matrices before multiplying by the coefficients and summing.

Note that it is possible to break the total energy of a conformation down into a set of terms which include the internal and backbone interaction energy of the lexide (or lexide pair) in question plus half the interaction of this lexide with the other lexides in the molecule. If it can be shown that a given lexide (or lexide pair) only interacts significantly with a certain number of its neighbors, then these terms can be pre-calculated and stored for all possible base sequences and used to further speed up total energy calculations. Thus, if, for example, a given lexide only interacts significantly with two lexide pairs on either side (see results section), then it will only be necessary to calculate the energies of each lexide pair within the fragment studied for $4^5 = 1024$ sequences.

Even without this approximation, ADAPT allows that the energies for different (pure or mixed) base sequences can be calculated very rapidly. This also means that it is easy to obtain the analytical derivatives of the energy with respect to the base coefficients. However, it is necessary to respect the normalization condition for the coefficients of each lexide during minimization and the requirement that all $C_{ik} \geq 0$. This is done by creating a new set of variables V_{ik} which are used to calculate the lexide coefficients C_{ik} as $V_{ik}^2 / \Sigma_{k=1,4} V_{ik}^2$ and automatically respect both requirements. The derivatives used by the minimizer must therefore be converted from $\delta E/\delta C_{ik}$ to $\delta E/\delta V_{ik}$. It should be noted that the $\delta E/\delta V_{ik}$ derivatives become identically zero for any pure base sequence. In order to better characterize these stationary points on the energy surface, we have calculated the second derivatives of the energy. The resulting Hessian matrix can be diagonalized to obtain the eigenvalues whose signs define the presence of a minimum, a maximum or of saddle points in various dimensions.

In addition to carrying out energy minimization there are two other ways to study sequence effects with ADAPT. Both of these approaches are limited to pure base sequences but have important applications. The first involves a combinatorial scanning of all possible sequences. This simply means setting the lexide coefficients so as to generate in turn each of the 4^N sequences which can be built up from N base pairs. This number rises exponentially with N, being roughly 10^6 for 10 base pairs, 4.3 x 10^9 for 16 base pairs and 2.8 x 10^{14} for 24 base pairs. However, given the speed of the energy calculations in ADAPT, it is possible to compare all possible sequences up to roughly 16 base pairs in length in modest computer times (building and testing a million sequences takes less than 5 seconds on a 500 MHz PC workstation running Linux). Beyond this limit, it is still possible to find a limited number of the lowest (or highest) energy sequences by using the energy approximation described above and by breaking down the combinatorial problem into steps, making a full combinatorial solution for overlapping sequence blocks and then only using a percentage of the lowest (or highest) energy solutions of the blocks for testing possible sequences for the full fragment.

The second approach can be termed gene threading and consists of calculating the energy difference between the target and reference conformations for all positions along the experimentally determined sequence of a whole genome or genome fragment. In effect the experimental sequence is "threaded" through the test conformations of our DNA fragments. Once again, the speed of energy calculations in ADAPT make it possible to deal with sequences many millions of base pairs in length and to produce "spectra" of the corresponding energy variations.

RESULTS

We will now discuss the application of our procedure to finding the optimal sequences for binding TBP. The calculations have been performed with three pairs of target and reference conformations containing respectively 10, 16 and 24 base pairs (built from 'N' lexides). In each case the reference conformation is an energy minimized B-DNA, while the target conformation is a B-DNA deformed so as to contain a Human TBP binding site conformation [17] at its center (see methodology). The energy matrices generated using JUMNA for these three pairs of conformations are used for energy minimization, combinatorial scans and for several examples of gene threading.

(i) Energy Matrices

Before discussing the actual search for TBP binding sites, it is worth saying a few words about the contents of the energy matrices generated by JUMNA. We will ignore the diagonal terms which include interactions between given lexides and the phosphodiester backbones of DNA. The remaining off-diagonal terms concern interactions between the bases of two lexides (or, in the present case, lexide pairs). A rapid inspection of these terms show that the elements involving neighboring base pairs are typically around 10 Kcal/mol. These values drop to roughly 1 Kcal/mol for second nearest neighbors and to around 0.02 Kcal/mol for third nearest neighbors. Beyond this point, values decrease only very slowly. If we look at the energy range as a function of sequence within any of these elements, we find roughly 2 Kcal/mol of variation for nearest neighbors, 0.2 Kcal/mol for next nearest neighbors and

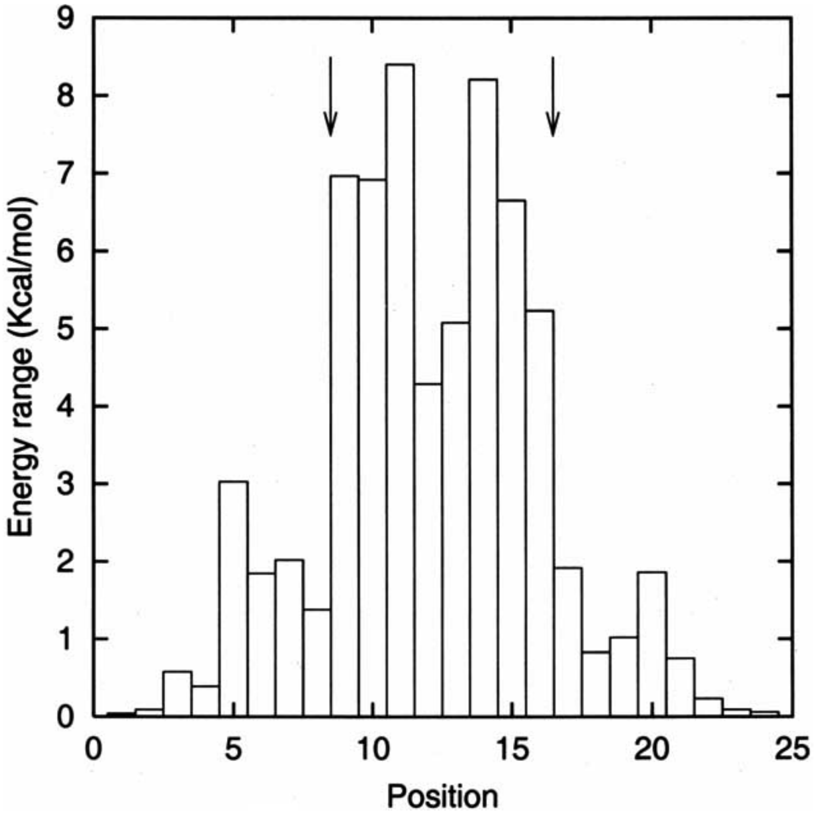

Fig. (1). Histogram of the sequence dependent energy range for the lexide pair deformation energies as a function of the lexide position within the fragment. Results refer to the 24 bp TBP target and B-DNA reference conformations. The position of the TBP binding site is delimited by the two arrows.

almost no variation for third neighbors. These results clearly imply that interactions beyond third nearest neighbor base pairs are of negligible importance for the energy model we are presently using.

It was mentioned in the methodology section that it is possible to divide the energy of a given DNA conformation containing N lexide pairs into a set of N lexide pair contributions. We can similarly sub-divide a deformation energy between the TBP target conformation and the B-DNA reference conformation into a set of lexide pair contributions, where each of these contributions takes into account the changes in the energy of the pair itself plus half the changes in its interaction energy with the rest of the DNA fragment. These contributions are naturally a function of the lexide coefficients. If we limit ourselves to pure sequences and, on the basis of the discussion above, to interactions with only two base pairs on either side of our chosen pair, then there are a total of $4^5 = 1024$ possible sequences to calculate.

We have made these calculations and plotted in Fig. (**1**) the overall range of each lexide pair energy for the 24 bp target and reference conformations. These results are interesting from several points of view. First, they clearly show that the 8 central lexide pairs are associated with the largest energy variations as a function of sequence. These pairs correspond to the position of the TBP binding site (delimited by the arrows shown in the figure) and they consequently undergo the largest changes in conformation between the reference and target conformations. Second, within the target site, we can see smaller variations for pairs 4,5 and 8 suggesting that they will play a lesser role in determining the best binding sequences. Third, if we sum the total sequence dependent energy variations for the 8 pairs of the binding site we obtain roughly 50 Kcal/mol, suggesting that this will be an upper limit for the energy variations which will be observed of any sequences tested against these target and reference conformations. We will see from the results below that these deductions are justified.

(ii) Energy Minimization and Combinatorial Scanning

It is first remarked that by using an average (N) lexide sequence for the target and reference conformations we can calculate a sequenced average DNA deformation energy for TBP binding. The values obtained for the 10, 16 and 24 base pair fragments are respectively 90.8, 99.1 and 111.7 Kcal/mol [Table (1)]. The presence of increasingly long B-DNA fragments on either side of the TBP binding site thus

Table 1. Results of Energy Minimization and of Combinatorial Searches for Optimal TBP Binding Sequences. Upper Case Letters within the Sequences Show the Location of the TBP Binding Site. All Energies are in Kcal/mol

Fragment	10 bp	16 bp	24 bp
Mean energy	90.8	99.1	111.7
Sequence	nNNNNNNNn	nnnnNNNNNNNNnnnn	nnnnnnnnNNNNNNNNnnnnnnnn
Energy minim.	69.3	73.0	84.7
Sequence	tTATTTAAAa	ccggTATTTAAAaacg	agcctcatTATTTAAAaacgacag
Global minim.	67.4	71.2	83.1
Sequence	tTATTTTTAa	ccggTATTTTTAaacg	agcctcatTATTTTTAaacgacag
Energy maxim.	112.5	125.1	138.0
Sequence	aGGGCCCTCc	gtcaGGGCCCTTctac	gtgagtcaGGGCCCTTctttcggt
Global maxim.	112.7	125.9	138.5
Sequence	aAAAGCCTCc	gacaAAAGCCTTctac	ctgagtcaAAAGCCTTctttcggt

increases the energy necessary for inducing the local protein binding conformation.

If we now carry out energy minimizations, we obtain the results shown in Table (1). Each of the three pairs of target and reference conformations lead to "pure" sequences in roughly 200-300 cycles of conjugate gradient minimization. The energy gain during minimization ranges from 21.5 Kcal/mol for the 10 bp fragments to 27 Kcal/mol for 24 bp fragment. Identical binding site sequences, TATTTAAA, are obtained with all three fragments. Diagonalization of the Hessian matrix for these sequences confirms that they are true minima with no negative eigenvalues.

It is also possible to find the worst possible sequence for each of our test fragments by simply reversing the order of

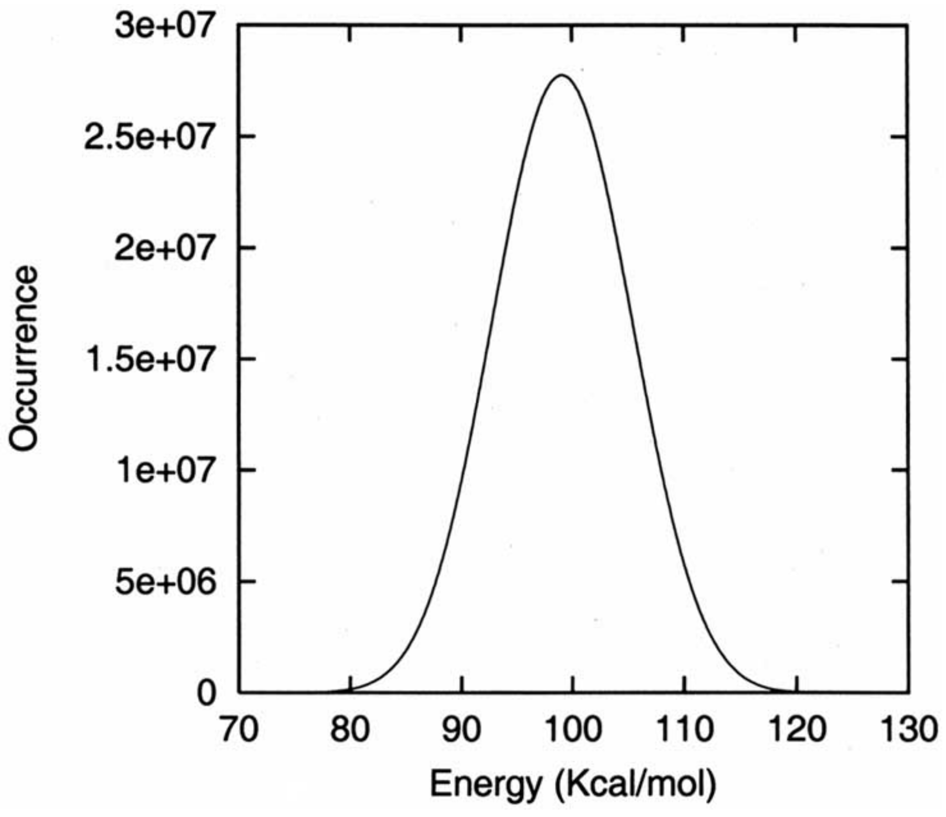

Fig. (2). Histogram of the deformation energy distribution obtained after a combinatorial search of all the 4.29 x 10^9 possible sequences for the 16 bp fragments.

the energy matrices passed to ADAPT (see methods sections). If we make this trial for our TBP targets, we again find a pure binding sequence which is either GGGCCCTC for the 10 bp fragment or GGGCCCTT for the 16 and 24 bp fragments. These results enable us to establish an overall energy range of roughly 40-50 Kcal/mol for the fragments studied as predicted by the study of the energy matrices in the preceding section. Given that the average energies of deformation to create the TBP binding site range from roughly 90-110 Kcal/mol, this implies that surrounding sequence effects alone can modify deformation energies by roughly ±25%.

In order to check the quality of the energy minimizations we have carried out combinatorial searches for each of the pairs of conformations we have used. This implies calculating the energies of 4^N sequences for N base pairs, or, respectively, 1.05 x 10^6, 4.29 x 10^9 and 2.81 x 10^{14} for the 10, 16 and 24 bp fragments we have created. For the first two fragments, we have made a full combinatorial search, but for the longest fragment, we used the simplifications described in the methods section to find only the 50 best and worst sequences. The results are given in Table (1). In each case, the combinatorial search has enabled us to find sequences that are more stable than those found by energy minimization. However, the energy gain is modest, ranging from 1.6 to 1.9 Kcal/mol for the three fragments tested. The global optimum binding sequence, TATTTTTA, is once again common to the 10, 16 and 24 bp fragments and it only differs from the sequence found by energy minimization by two T⇔A inversions in positions 6 and 7. The global maxima located, also listed in Table (1), all have energies slightly above those located by energy maximization, but the difference is again small (less than 1 Kcal/mol). The corresponding binding sequences show more variation with respect to those found by energy maximization, AAAGCCT(T/C) versus GGGCCCT(T/C). We also remark that the overall distribution of energy minima closely follow a Gaussian distribution, as shown in Fig. (**2**) for the 16 bp fragments and that the total range of deformation energies as a function of sequence is 45.3 Kcal/mol for the 10 bp fragments, 54.7 Kcal/mol for the 16 bp fragments and 55.4 Kcal/mol for the 24 bp fragments. These values are again in good agreement with the predictions made on the basis of the energy matrices in section (i).

The 50 best sequences found by combinatorial searching can be used to learn more about the nature of the energy surface we are studying (namely an energy hypersurface in sequence space). We will consider the results obtained with the 10 bp fragments. By calculating and diagonalizing the Hessian matrices for these sequences we find that only 3 out of the 50 best sequences actually correspond to energy minima (that is, with no negative eigenvalues). These true minima are respectively, the 1st (tTATTTTTAa, 67.4 Kcal/mol), the 23rd (tTATAATTAa, 69.0 Kcal/mol) and the 33rd (tTATTTAAAa, 69.3 Kcal/mol). We have already discussed two of these sequences since the 1st sequence naturally corresponds to the global energy minimum found by combinatorial searching and the 33rd sequence corresponds to the minimum found by energy minimization. The third true minimum is again close to the globally optimal sequence (differing by two T⇔A inversions, this time in positions 4 and 5) and to the globally optimal energy. All the remaining 50 best sequences have at least one negative eigenvalue. The number of negative eigenvalues tends to increase as the energy of the corresponding sequence increases and several sequences towards the end of our list have 5 or 6 negative eigenvalues. This trend is confirmed if we look at the 50 worst sequences, where all members of the list have at least 26 negative eigenvalues.

These results explain why energy minimization is successful in locating very low energy sequences, even if it

Table 2. Comparison of the Experimental Consensus Sequence of Human TBP [22] and of the Consensus Derived from Combinatorial Calculations Using the 10 bp TBP Target and Reference Fragments and an Energy Cutoff of 5 Kcal/mol with Respect to the Global Minimum. Consensus Base Codes are as Follows : K = G or T, S = G or C, R = A or G, W = A or T and N is a Non-specific Base

	TRANSFAC (%)					ADAPT (%)				
Pos	A	C	G	T	Code	A	C	G	T	Code
1	16	37	39	8	S	15	20	30	35	K
2	4	12	5	79	T	0	4	1	95	T
3	91	0	0	9	A	97	0	0	3	A
4	1	3	0	96	T	0	1	0	99	T
5	91	0	1	8	A	32	8	8	52	W
6	69	0	0	31	A	39	0	3	58	W
7	92	1	5	2	A	43	0	1	57	W
8	57	1	11	31	W	43	1	0	56	W
9	40	11	40	9	R	52	6	35	8	R
10	14	35	39	12	N	32	20	22	26	N

Table 3. TBP Binding Sites from the TRANSFAC Database [22] used for Tests of Gene Threading. The Sites are Identified by their TRANSFAC Accession Number and by their Position Within the Sequence. 24 bp Long Sequences are Shown in Each Case. The Bases Belonging to the TRANSFAC Site are Underlined. The Bases Belonging to the ADAPT Threading Site Located are Shown in Capitals and the Start of this Site is Listed

Site	Acc. no.	Binding sequence	Position	Start
HSP	R00770	tgacgactTATAAAAGcccaggg	246:252	246
DHFR	R03158	ctc gcctgCACAAATAgggacgag	289:308	294
GFAP	R03167	cccactccTTCATAAAgccctcgc	2133:2140	2133
HSU6RNA	R03171	ttcttggcTTTATATAtcttgtgg	232:246	235
PF4	R01218	gcagtgaaGATAAAACgtgtctag	357:387	370

does not find the global minima. Although we are working in relatively high dimensional spaces, with respectively 30, 48 and 72 variables for our 10, 16 and 24 base pair fragments, the study of the Hessian matrices suggests that there are very few low energy minima and that the surface is thus strongly funneling.

If we group together the best energy solutions, we can generate what amounts to a consensus sequence for our energy matrices. We have done this for the 10 bp fragment using all energies within a somewhat modest limit of 5 Kcal/mol from the global optimum energy of 67.4 Kcal/mol. The results are shown in Table (2). Given that we have only taken into account the deformation caused by TBP and not the specific TBP-DNA interactions our results are encouragingly close to the experimental consensus. Our approach yields kTATWWWWRn compared for Human TBP binding consensus of sTATAAAWRn given in the TRANSFAC data base [22] under the accession number M00252 (note K = G or T, S = G or C, R = A or G, W = A or T and N is a non-specific base). It is interesting to note that our major failing is not identifying clear cut A's in positions 5-7 and it is indeed at the first two of these positions that specific amino acid - base hydrogen bonds occur in the crystallographic TBP-DNA complex.

(iii) Gene Threading

Searching for TBP binding sites within genome sequences has been tested on 5 of the 23 binding sites for human TBP currently listed in the TRANSFAC database [22]. These test sites are listed in Table (3) and involve four human genes, heat shock protein (HSP), dihydrofolate reductase (DHFR), glial fibrillary acidic protein (GFAP), human sub-unit 6 small nuclear RNA (HSU6RNA) and one rat gene, platelet factor 4 (PF4). The corresponding sequences were retrieved from the database and used for threading against the TBP target and reference conformations with 10, 16 and 24 base pairs described above.

The overall results for the first gene, HSP, are shown in Fig. (**3**) for the 10 bp test conformations. This gene sequence contains 2691 base pairs. As it is threaded through the target and reference conformations, the energy necessary to create the TBP binding conformation varies by roughly ±20 Kcal/mol around an average value of roughly 90 Kcal/mol. This is perfectly compatible with the results found by combinatorial searching. The energy "spectrum" in Fig. 3 shows rapid fluctuations which sharply define sequences strongly favoring or strongly disfavoring binding. These variations are explained by the fact that TBP binding requires an important DNA deformation which opens the minor groove and bends the molecule in the direction of the major groove. A sequence which facilitates this change will produce a low deformation energy when is perfectly positioned with respect to the target conformation, but will be likely to produce a very large deformation energy when it is offset by a few base pairs and falls out of helical phase with the deformation it facilitates.

The lowest energy along this sequence falls at position 246, which, in this case, is the experimental HSP binding

Table 4. Optimal Binding Sites Found by ADAPT Gene Threading Using a TBP Target and a B-DNA Reference Conformation. The Three Results for Each Sequence, which have been Aligned at the Start of the TBP Binding Site, Correspond to Tests with 10, 16 and 24 bp Fragments

Site	Optimal ADAPT sequence	Position
HSP	tTATAAAAGc	246
	ccatTTTTTAAGttgg	2629
	caggccatTTTTTAAGttggttac	2629
DHFR	tTATTAAAAa	1064
	ccatTATTAAAAaatt	1064
	tttaccatTATTAAAAaatttttg	1064
GFAP	tTTTTTTTGt	1664
	ccttTTTAATTGatgc	1827
	ttttttttTTTTTTTGtgagacaa	1664
HSU6RNA	aTATTTTTAc	366
	tccaTATTTTTAcatc	366
	gcgttccaTATTTTTAcatcaggt	366
PF4	tTATTTAATt	1486
	acatTATTTTGAaggg	1404
	tacctctgTATAAGAAaataatca	1147

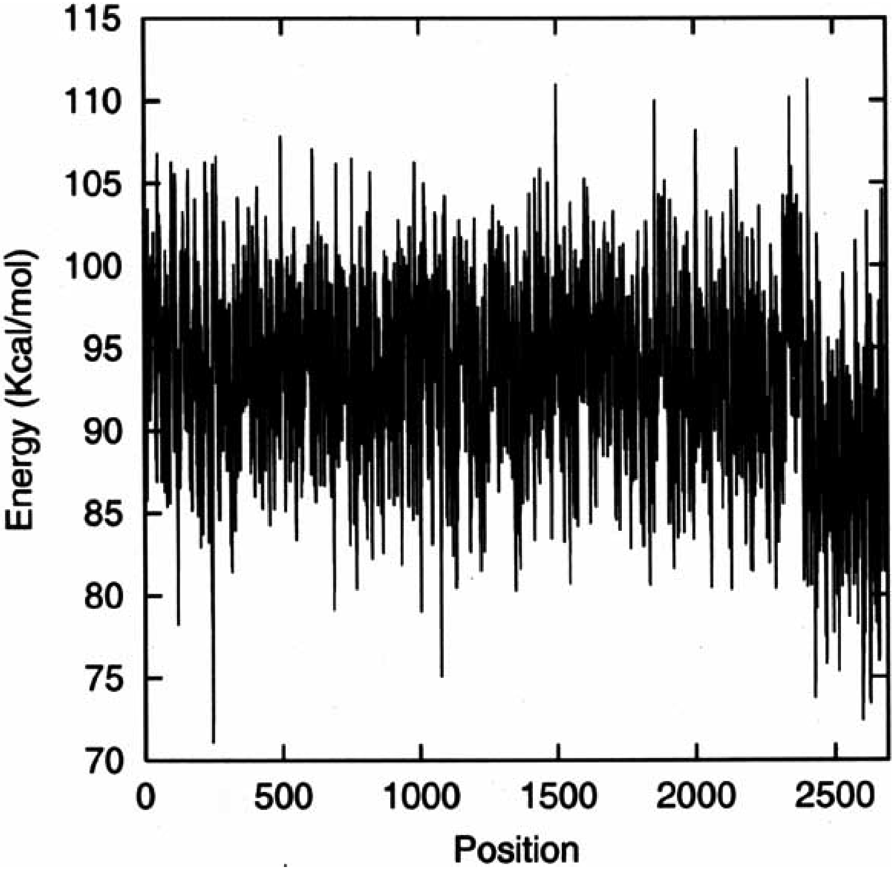

Fig. (3). Deformation energy spectrum obtained by threading the 2619 base pairs of the HSP gene sequence through the 10 bp TBP target and reference fragments.

site shown in Table (4). The energy spectrum in this region can be seen in more detail in Fig. (**4a**). This figure also compares the results obtained with the 16 and 24 bp test fragments. These results show that the binding site falls exactly at position 246 with all three fragments. Indeed, there is very little variation between the corresponding energy spectra, beyond an overall shift to higher deformation energies as the test fragments become longer. These shifts of roughly 10 Kcal/mol in each case are again in line with the change of average TBP binding energies discussed above for the different test fragments.

Although the true binding site corresponds to the global minimum in the 10 bp energy spectrum, there are other favorable sites with very similar energies. Two examples of this are given in Fig. (**4b**), which shows a minimum at 1077 and Fig. (**4c**) which shows two minima at 2601 and 2629. In fact, the latter region contains the global minima for

Table 5. ADAPT Gene Threading for Each of the TBP Sites Listed in Table (3). The Position of the Experimental Binding Site within the Energy Spectrum from ADAPT is Characterized by the Ranking of the Site in Terms of the Energy Minima Located, the Percent of Sites at or Below this Energy and the Percentage this Energy Represents in Terms of the Full Range of Energy Variation Observed (Roughly 40 Kcal/mol). The Last Line of the Table Refers to a Mutated PF4 Binding Site (See Text for Details)

Site	Length	Rank (10bp)	Rank (16bp)	Rank (24bp)	% Sites	% Energy
HSP	2619	1 st	4 th	3 rd	0.1	3.7
DHFR	1275	23 rd	6 th	11 th	1.0	21.7
GFAP	2630	36 th	54 th	48 th	1.7	20.9
HSU6RNA	464	28 th	33 rd	27 th	6.3	21.8
PF4 (GATA)	1675	28 th	13 th	15 th	1.1	14.9
PF4 (TATA)	1675	3 rd	3 rd	2 nd	0.2	2.9

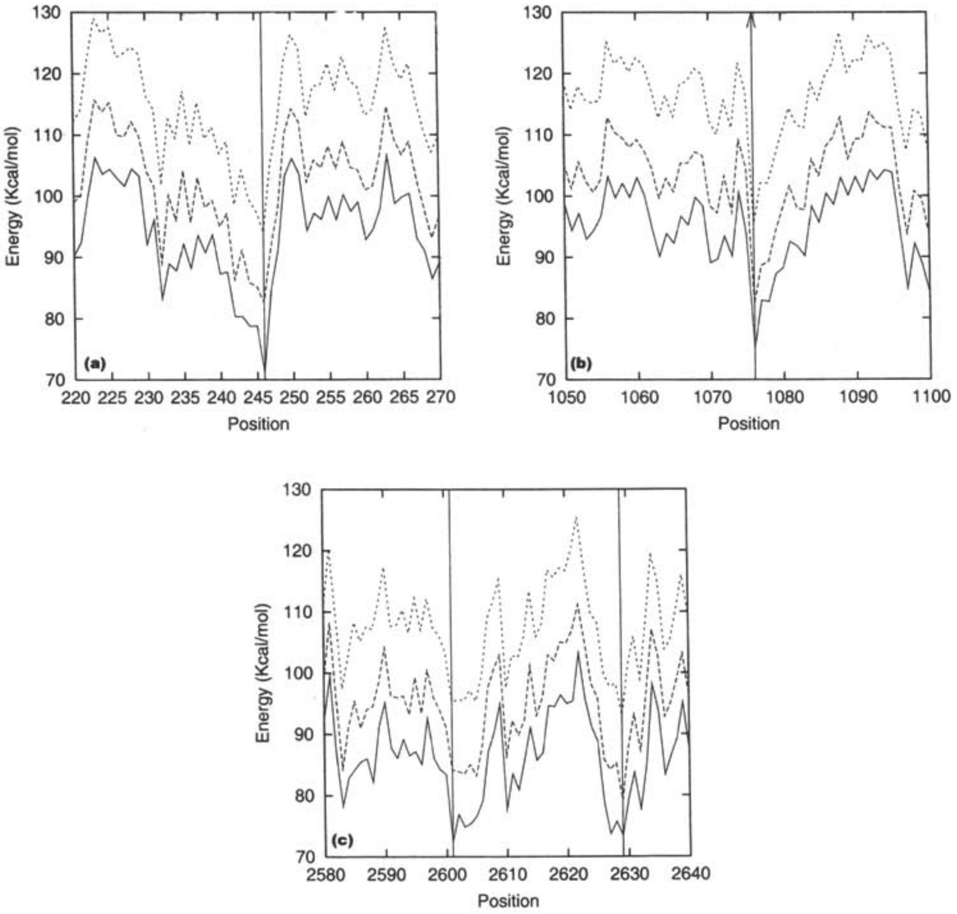

Fig. (4). Details of the HSP energy spectra obtained with the 10 bp, 16 bp and 24 bp TBP target and reference fragments (increasing average energies) : (a) the experimental binding site at position 246 (indicated by a vertical line), (b) a secondary site at position 1076, (c) two further secondary sites at positions 2601 and 2629. The latter being the global energy minimum for the 16 and 24 bp fragments.

the 16 and 24 bp fragments, which both fall at position 2629. The corresponding sequences are listed in Table (4).

Similar tests were carried out for the DHFR, GFAP, HSU6RNA and PF4 sites. The results are presented in Tables 4 and 5. The energy spectra obtained from the sequences containing these sites look very much like those shown for HSP. They cover a very similar energy range and also show the same increasing average energy values for the 10, 16 and 24 bp test fragments. None of these sites correspond to the global minima of the energy spectra, but they all fall in the low energy range of each spectrum. The details are given in Table (5). We are searching for sites that typically fall within the 20-30 best minima of a given energy spectrum. Since, with the exception of HSU6RNA, the sequences studied contain 1500-2500 bp, this places the sites within the best 1%-2% of all possible binding positions. This corresponds to roughly the best 20% in terms of deformation energy.

It is also interesting to note that, with the exception of HSP, the sites we seek are rather far from the normal consensus target of TBP (compare the data in Table (2) and Table (4)). Despite this fact, they fall within the best few percent of the energy spectra. Not surprisingly, the optimal sites found by threading, listed in Table (5), are generally closer to the TBP consensus.

Finally, an experimental study of the PF4 site [23] gives us the opportunity to look at the impact of a sequence mutation of gene threading. The PF4 site contains the sequence GATAAAA [370:376]. TBP binding to this site is inhibited by preferential binding of the GATA-1 factor. This situation can be reversed by a G to T mutation at position 370. The results of this mutation on the energy spectrum for the 10 bp TBP test fragments are shown in Fig. (**5**). Changing G to T has a strong impact on the spectrum for positions at or just before the mutation. It notably lowers the binding energy for the minimum at position 370 by almost 5 Kcal/mol and changes this site from 28th to 3rd in the energy ranking. Similar results are obtained with the 16 and 24 bp fragments as shown in Table (5).

CONCLUSIONS

We have developed an original method for analyzing genome sequences on the basis of the physical properties of

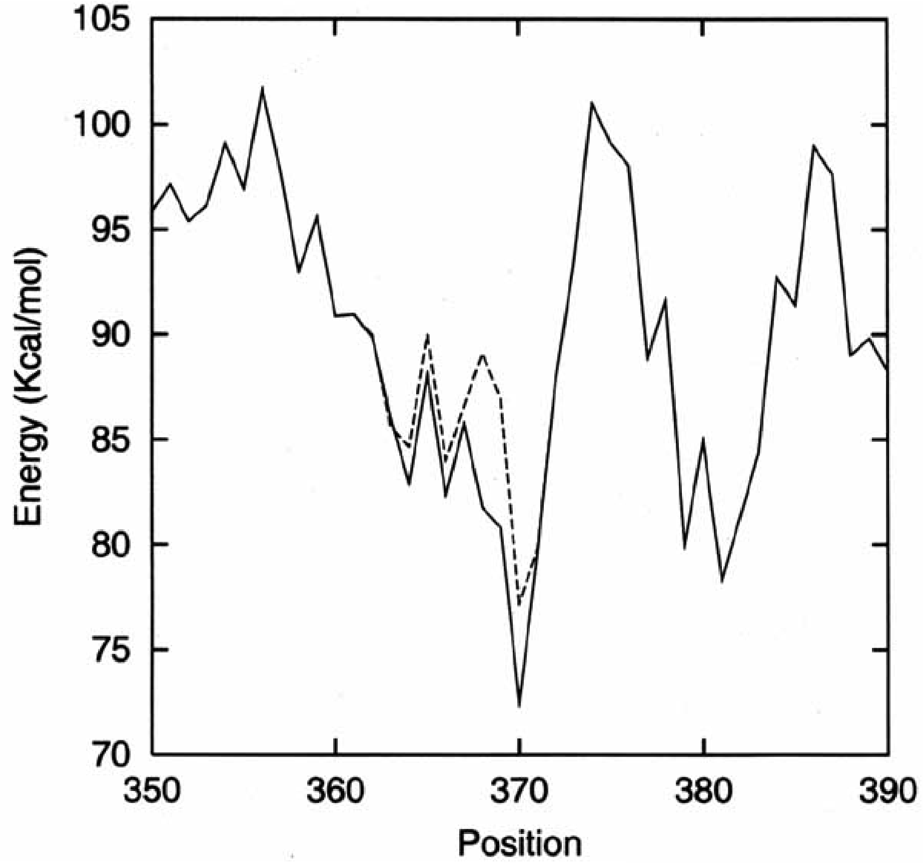

Fig. (5). Energy spectrum for the PF4 binding site before (dotted line) and after (solid line) mutation of the binding sequence at position 370 from <u>G</u>ATAAAA to <u>T</u>ATAAAA.

the corresponding DNA. This approach, which combines a modified JUMNA program with a new program called ADAPT, is fast but conserves an all-atom model for the determination of the molecular properties. Given the energy analysis we have carried out, our approach can be viewed as an overlapping pentanucleotide model, but one in which the parameters describing each pentanucleotide are dependent on the position of the pentamer within the binding site and also on the nature of the binding site. These features make this approach much more detailed than current methods based on fixed tables of di- or trinucleotide properties.

For a given DNA deformation, this approach can be used to locate the sequence which will minimize the corresponding deformation energy. This can be done with certainty by combinatorially searching all possible sequences. Although simple energy minimization does not currently locate the global minimum, the nature of the sequence space hypersurface presented here suggests that a minima-hopping optimization algorithm would probably have a good chance of succeeding and such tests are underway.

We have also introduced a gene threading approach to search for binding sites within very long sequences. The results obtained for TBP using this method are encouraging and suggest that known binding sites fall within the best 1%-2% of possible sites, even when they are relatively far from the canonical consensus sequence of the protein.

It should be added in conclusion that this method can probably be improved by taking into account protein-DNA interactions, and we are working on this extension. It might also be necessary to consider other force fields or solvent representations. Happily, these changes in no way represent handicaps since the duration and complexity of the method used to generate the energy matrices for ADAPT have no impact on the speed of subsequently searching for potential binding sites.

ACKNOWLEDGEMENT

The authors wish to thank IFCPAR for their support of this work as part of the project 1804-1 on protein-DNA interactions and would also like to thank K. Zakrzewska for help with the study of the energy hypersurface carried out in this article.

ABBREVIATIONS

DHFR	=	Dihydrofolate reductase
GFAP	=	Glial fibrillary acidic protein

HSP	=	Heat-shock protein
HSU6RNA	=	Human U6 small nuclear RNA
JUMNA	=	Junction minimisation of nucleic acids
PF4	=	Platelet factor 4
TBP	=	Tata-box binding protein

REFERENCES

[1] Karas, H.; Knüppel, R.; Schultz, W.; Sklenar, H.; Wingender, E. *CABIOS* **1996**, *12*, 441.

[2] Pedersen, A.G.; Baldi, P.; Chauvin, Y., Brunak, S. *Comput. Chem.* **1999**, *23*, 191.

[3] Beveridge, D.L.; McConnell, K.J. *Curr. Opin. Struct. Biol.* **2000**, *10*, 182.

[4] Bruckner, I.; Sánchez, R.; Suck, D.; Pongor, S. *EMBO J.* **1995**, *14*, 1812.

[5] Lafontaine, I.; Lavery, R. *Biophys. J.* **2000**, *79*, 680.

[6] Burley, S.K. *Curr. Opin. Struct. Biol.* **1996**, *6*, 69.

[7] Wobbe, C.R.; Struhl, K. *Mol. Cell. Biol.* **1990**, *10*, 3859.

[8] Starr, D.B.; Hoopes, B.C.; Hawley, D.K. *J. Mol. Biol.* **1995**, *250*, 454.

[9] Parvin, J.D.; McCormick, R.J.; Sharp, P.A.; Fisher, D.E. *Nature* **1995**, *373*, 724.

[10] Flatters, D.; Lavery, R. *Biophys. J.* **1998**, *75*, 372.

[11] de Souza, O.N.; Ornstein, R.L. *Biopolymers* **1998**, *46*, 403.

[12] Pastor, N.; Pardo, L.; Weinstein, H. in *Molecular Modeling of Nucleic Acids*, A.C.S. Symposium series 682 eds. Leontis, N.B.; SantaLucia Jr., J **1998**, p329.

[13] Pastor, N.; Pardo, L.; Weinstein, H. *Biophys. J.* **1997**, *73*, 640.

[14] Kim, Y.; Geiger, J.H.; Hahn, S.; Sigler, P.B. *Nature* **1993**, *365*, 512.

[15] Kim, J.L.; Nikolov, D.B.; Burley, S.K. *Nature* **1993**, *365*, 520.

[16] Juo, Z.S.; Chiu, T.K.; Leiberman, P.M.; Baikalov, I.; Berk, A.J.; Dickerson, R.E. *J. Mol. Biol.* **1996**, *261*, 239.

[17] Nikolov, D.B.; Chen, H.; Halay, E.D.; Hoffman, A.; Roeder, R.G.; Burley S.K. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 4862.

[18] Marchler-Bauer, A.; Byrant, S.H. *Trends in Biol. Sci.* **1997**, *22*, 236.

[19] Lavery, R. in "*Structure and Expression Vol.3 DNA Bending and Curvature*" eds. Olson, W.K.; Sarma, R.H.; Sarma, M.H.; Sundaralingam M. Adenine Press **1988**, 191.

[20] Lavery, R.; Zakrzewska, K.; Sklenar, H. *Comp. Phys. Commun.* **1995**, *91*, 135.

[21] Lebrun, A.; Lavery, R.; Weinstein, H. *Protein Engineering* **2001**, *14*, 233.

[22] Wingender, E.; Dietze, P.; Karas, H.; Knippel, R. *Nucleic Acids Res.* **1996**, *24*, 238. http://transfac.gbf-braunschweig.de

[23] Aird, W.C.; Parvin, J.D.; Sharp, P.A.; Rosenberg, R.D. *J. Biol. Chem.* **1994**, *269*, 883.

Received: 28 August 2000 **Accepted: 8 January 2001**