Word Design for Biomolecular Information Processing

J. Ackermann and F.-U. Gast[a]

Fraunhofer Gesellschaft, Schloss Birlinghoven, D-53754 Sankt Augustin
[a] Justus-Liebig-Universität Gießen, Institut für Anorganische und Analytische Chemie,
Schubertstraße 60, Haus 16, D-35392 Gießen

Reprint requests to Dr. J. A.; Fax: -141511, E-mail: joerg.ackermann@gmd.de

Z. Naturforsch. **58a,** 157 – 161 (2003); received October 10, 2002

The design of DNA sequences plays a fundamental role for many biomolecular applications and is one of the most important theoretical tasks to fathom the potential of molecular information processing. Optimization strategies have been based on the model of stiff "digital" polymers by counting the number of base mismatches (Hamming distance and related distances). In this work we show the limitation of such a combinatorial approach because of the ability of DNA to build more complex structures. We develop a model platform to optimize word sets according to all possible secondary structures occurring for the relevant word-word interactions. The fidelity of the hybridization reactions can be improved significantly and as an example of a set of 24 words of 16-mers we show that the optimal set has unique physical properties, such as binding energy, melting temperature, and G+C content.

Key words: DNA Library; DNA Computing; Hybridization; Folding; Partition Function.

In nature, the genesis, evolution, and existence of living systems is based on complex biomolecular functions. In these processes, biopolymer strands play the role of information media carrying – for example – the construction plan of entire organisms. Adleman [1] has demonstrated the feasibility of biomolecular DNA processing for solving combinatorial problems. In this way he added a new point of view to the world of biomolecular processes with terms like "biomolecular information processing" or "DNA computing". Thereupon the undreamed-of possibilities of biomolecular information processing have been discussed and tested by an increasing community of scientific groups, see [2 – 8] for a selection of papers.

The present work deals with the task to find a set of sequences of given length (words) that are unique in the sense that their hybridization properties are well distinguishable from each other. This means that each "word" of the set should bind specifically to its Watson-Crick complement and not to any other member of the set nor their Watson-Crick complements. The design process is demonstrated for the goal to encode binary information by concatenations of DNA words. For each bit we need two DNA words, one for the "0" and one for the "1". To code a bit string of length N we need $2 \times N$ different DNA words: $V_i^0, V_i^1, i = 0, 1, 2, \ldots, N-1$. The upper index indicates the value of the bit (either zero or one) and the lower index gives the bit order. A population of such assembled strands encoding, for example, all possible solutions of a combinatorial problem, can be produced by standard biomolecular techniques. Examples of such a combinatorial problem are the Hamiltonian path problem [1], the maximal clique problem [4], and the satisfiability problem (SAT) [7 – 9], *e.g.,* the "Knight" chess problem [6].

Those and only those strands containing a one (or zero) at a given position in the binary string must be separable by a specific hybridization step, which can be realized by using Watson–Crick complements of certain words immobilized on beads [1], on surfaces [7], or in gels [9]. A false positive selection will lead to a wrong or at least to a statistically noisy result.

The task to find an optimal set of words is difficult for two reasons. Firstly, the number of possible sets of words is incredibly high. For an experimentally reasonable word length of sixteen nucleotides one has $4^{16} = 2^{32} \approx 4 \times 10^9$ different words. The number of different sets containing only 30 words is higher than 10^{243}, and a straightforward test of all sets is not practical. Secondly, the binding of two strands is a complicated dynamic process of folding and unfolding in

0932–0784 / 03 / 0200–0157 $ 06.00 © 2003 Verlag der Zeitschrift für Naturforschung, Tübingen · http://znaturforsch.com

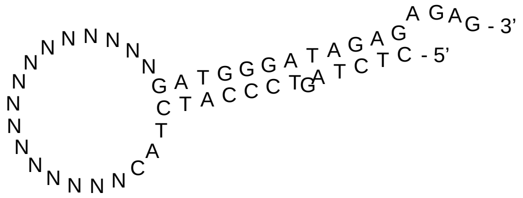

Fig. 1. Example of a secondary structure describing a mismatch binding (see text).

three-dimensional space, and thus the theoretical quality criteria have to compromise on the level of approximation.

Despite the fact that the fidelity of the hybridization reaction is the limiting factor for the scalability of DNA computing to larger problems [9], word design strategies applied so far were mainly based on combinatorial constraints like the Hamming distance and/or related distances [10]. A more reliable measure for the relative stability of a DNA duplex structure is its free energy [8]. The free energy of perfectly matching DNA duplex structures can be easily computed by applying the nearest neighbor approximation and thermodynamic parameters derived from melting experiments [11]. For structures with mismatches no general model for predicting the free energy is available. So-called staggered zipper models based on thermodynamic data turned out to be inadequate, because they neglect configurations containing internal loops, hairpins, bulges, as well as single and tandem mismatches [12].

How to deal with such structures is well-known in the field of RNA/DNA secondary structure prediction [13], and the thermodynamic stability of a DNA duplex structure is closely related to the thermodynamic stability of a corresponding hairpin structure [14]. Hence, to describe the hybridization of two strands (word$_i$ and word$_j$) we have constructed sequences where both strands are connected by a spacer sequence:

$$5'\text{–word}_i\text{\_spacer\_word}_j\text{–}3'.$$

The spacer sequence consists of artificial nucleotides (denoted by "N") which are defined to have no physical binding properties. No binding is possible to the nucleotides in this region. The secondary structure of such a DNA sequence can be computed by applying a (dynamic programming) folding algorithm (i.e. the Vienna folding package [13]) and appropriate DNA energy parameters [11]. A hairpin describing a duplex structure with mismatches is shown in Figure 1. Note that the bulge shown in Fig. 1 cannot be avoided by any combinatorial constraint. Physically the spacer confines the relative motion of two binding partners and thus defines an effective concentration. The entropy contribution to closure the hairpin loop formation corresponds to the concentration dependend entropy increment in the case of a second order hybridization reaction. For a hairpin loop of length $l = 16$, the corresponding effective total strand concentration C can be estimated from the free energy contribution ΔG_{loop} by $C = 4\exp(\Delta G_{\text{loop}}/RT) \approx 1$ mM (for $\Delta G_{\text{loop}} = 5$ kcal/mol at the temperature $T = 37$ °C; R denotes the molar gas constant). In the difference of the free energies for hairpin structures of identical length this contribution cancels. Small variations in the length of the various hairpin loops (see Fig. 1) lead to free energy differences

$$\Delta G_{\text{loop}}(l+\Delta l) - \Delta G_{\text{loop}}(l) = \frac{3}{2}\, RT \ln(1+\Delta l/l),$$

which becomes negligible for a minimum loop length of $l = 16$. A correction of this effect is easily possible by changing the energy parameters or by a minor modification of the algorithm. This, however, would not influence the results presented here (results not shown), but would make a comparison to energies obtained with other folding programs difficult.

Since we are not interested in the stability of a certain duplex structure, but in the overall thermodynamic stability of a (mismatch) word-word interaction, we calculate the partition function Q of all possible duplex structures. Therefore the Vienna RNA folding package [13] with appropriate DNA energy parameters [11] was applied. Strongly paired bases result from the matrix of base pair probabilities (derived via backtracking from the partition function), and these pairings describe a folding structure (see Fig. 1).

The free energy $\Delta G = -RT\ln(Q)$ is the most direct quantity to characterize the strength of a binding. A standard value of $T = 37$ °C is chosen for all results presented here. The free energy ΔG corresponds to an effective equilibrium constant for an ensemble of probability weighted duplex structures and has to be distinguished from the minimal free energy, which describes the hybridization for one particular structure (i.e. the minimal free energy structure).

The alternative to define and use appropriate sums of base pairing probabilities is not considered in this work. Whereas the correct binding leads to one characteristic binding energy denoted by ΔG_{B}, the possible mismatch pairings of a word (binding to other words in the set, and bindings to complements of other words in the set) produce spectra of free energies. The strongest mismatch binding of a word corresponds to the lowest free energy in these spectra, in the following denoted ΔG_{I}. A set of words is characterized by its spectrum Σ_{B} of binding energies (ΔG_{B}) and the spectrum Σ_{I} of lowest mismatch binding energies (ΔG_{I}). The main quality criterium applied here is the energy gap between Σ_{B} and Σ_{I} defined by $\delta F = \min(\Delta G_{\mathrm{I}} - \Delta G_{\mathrm{B}})$ for all $\Delta G_{\mathrm{B}} \in \Sigma_{\mathrm{B}}$, $\Delta G_{\mathrm{I}} \in \Sigma_{\mathrm{I}}$.

Stochastic search algorithms have been used successfully for decades in the construction of good binary codes. We found the following simple random search algorithm preferable to maximize the energy gap δF for a given set of words:

1. Calculate Σ_{B}, Σ_{I} and δF for the set; save δF.
2. Select a word w_i randomly.
3. Construct a random DNA sequence w_{random}.
4. Recalculate Σ_{B}, Σ_{I} and δF_{new} for the set but with w_i replaced by w_{random}.
5. In the case of $\delta F_{\mathrm{new}} \geq \delta F$ accept the replacement $w_i = w_{\mathrm{random}}$ and go back to step 1. Otherwise go back to step 3.

Note that the condition $\delta F_{\mathrm{new}} \geq \delta F$ enables a replacement of nearly all sequences even when the gap δF has reached its maximum value. This introduces, similar to a simulated annealing algorithm, an effective noise term which is sufficient to exploit the enormous search space. Starting with the set of 24 words optimized according to combinatorial constraints [15] we initially obtain a Σ_{B} in the range of $\Delta G_{\mathrm{B}} = -17.9$ kcal/mol to $\Delta G_{\mathrm{B}} = -14.2$ kcal/mol, and Σ_{I} in the range of $\Delta G_{\mathrm{I}} = -7.0$ kcal/mol to $\Delta G_{\mathrm{I}} = -3.1$ kcal/mol. Hence the binding energies of correct and incorrect binding are separated by a minimal gap of $\delta F = 7.2$ kcal/mol.

Applying the random search algorithm described above, the mismatch binding energies ΔG_{I} increase to values in the range of $\Delta G_{\mathrm{I}} = -4.4$ kcal/mol to $\Delta G_{\mathrm{I}} = -3.9$ kcal/mol, whereas Σ_{B} converges to values in the range of $\Delta G_{\mathrm{B}} = -19.6$ kcal/mol to $\Delta G_{\mathrm{B}} = -19.1$ kcal/mol (see Table 1 for the sequences). Thus the energy gap δF has more than doubled, from initially 7.2 kcal/mol to 14.7 kcal/mol. The melting temperatures of the words increase by approximately 10 °C to an average value of 74.4 °C ± 0.5 °C with a narrow total range from 73.4 °C to 75.6 °C (at 1 M salt, 5 μM strand concentration). These numbers may be compared with the properties of the word set applied recently to solve a nontrivial 20 variable 3SAT problem [9]. The thermodynamic discrimination between correct and mismatch binding in their set correspond to an energy gap of $\delta F = 4.1$ kcal/mol (computed by the method described above), which is more than three times lower than the value obtained for the set in Table 1.

Table 1. Set of twenty four DNA sequences (written in 5' to 3' direction), optimized to discriminate wrong selection in biomolecular computing.

i	V_i^1	V_i^0
0	CGCAA GGCTA ACCCC G	ACACG AGCAC GATGC C
1	GCTCA CCGCG ATTCC A	CGTCT GTCCT GCACC G
2	CCACG TCGTT CGTCC C	GCTTG CTTGC CACCC T
3	TCCCC CTCCC GATCG A	AGCGG ACCAA TGCCA C
4	GCGTG TGGGA TCTCG C	GTACC AGTCG CAGCG C
5	CGGAG AAACA GCGGC C	CGCTC CTTCG CACTG T
6	GCACA CACCC TCGAC G	GGCGG GTCGA GAATC G
7	GTGAG ACGCT GGCAG G	TTGCT ACCTC GGGGC G
8	CGCTG AAGAG GCCGA G	TGGCA GCCCA TTGTC G
9	GCGCA TCTCC CAGAG C	GCCGA TCCTA GCCGG A
10	CCCAA GCGTG ACAGG C	CGTGA GCTTC CGACC G
11	AGGGC GCTTT GGATG C	TGGTC CCAAC TGGCG T

The final physical properties of the set in Table 1 seem to be rather arbitrary. In order to study whether an optimal set of words can also be optimized for a completely different binding energy range, we changed the random search algorithm in the following way:

a) maximize the mismatch binding energies ΔG_{I} with the constraint that all binding energies ΔG_{B} are below certain given upper bound values.

b) minimize the binding energies ΔG_{B} with the constraint that all mismatch binding energies ΔG_{I} are above certain given lower bound values.

The spectra Σ_{B} and Σ_{I} resulting for various upper (word set number 1 – 7) or lower bounds (word set number 9 – 14) are plotted in Figure 2. The word set for a maximal gap δF is located between them (word set number 8). For each a word set the binding energies ΔG_{B} (triangles) and the mismatch binding energies ΔG_{I} (diamonds) are aligned in vertical direction. The lower bounds for ΔG_{I} are drawn as dotted (upper left) diagonal line and the upper bound for ΔG_{B} is given by the a hatched (lower right) diagonal line.

Relaxing the lower bound for the mismatch binding energies ΔG_{I} has a direct effect on the properties

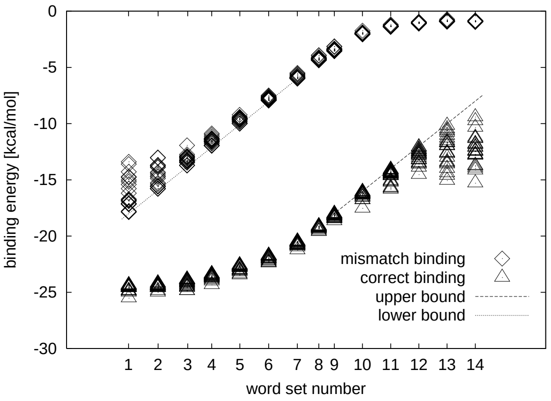

Fig. 2. Mismatch binding energies (ΔG_{I}, diamonds) and binding energies (ΔG_{B}, triangles) for various optimized word sets, see text.

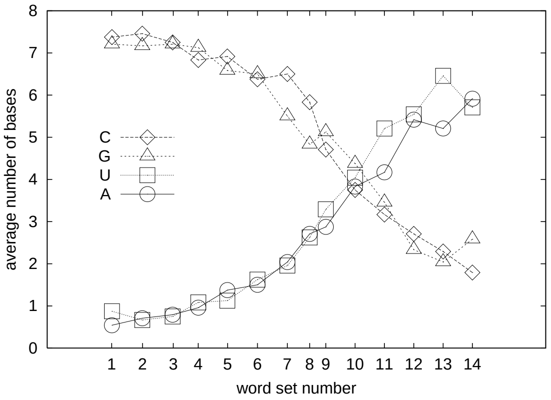

Fig. 3. Average base content of the optimized word sets.

of the resulting optimized word sets. The tolerance in the mismatch binding is exploited to improve the strength of the correct binding; the optimized energies ΔG_{B} become more negative after decreasing the lower bound. The ability to improve the binding strength is, of course, limited to a certain amount and is accompanied by a higher G+C content and a higher melting temperature (see Fig. 3 and Fig. 4, respectively). Increasing the G+C content, the energies ΔG_{I} follow their lower bound, but simultaneously the range of Σ_{I} becomes broader indicating that the lower bound for ΔG_{I} is no longer the limiting criterion for the selection of the word sequences. The mismatch binding energies ΔG_{I} can reach similar values as the binding energies ΔG_{B}, and thus the energy gap δF finally decreases.

In contrast, relaxing the upper bound for the binding energies ΔG_{B} changes the properties of the word set in the opposite direction. The binding energies ΔG_{B} follow their upper bound, accompanied by a broadening of the range of the values showing that the upper bound for ΔG_{B} is no longer the limiting criterion for the selection of the word sequences. This behavior is correlated to the decline of the melting temperature and the broadening of the range of T_{m} values for increasing upper bounds (see Figure 4). The freedom to use sequences with low melting temperatures is exploited to minimize the strength of the mismatch bindings. The increase of ΔG_{I} is also accomplished by lowering the G+C content (see Figure 3).

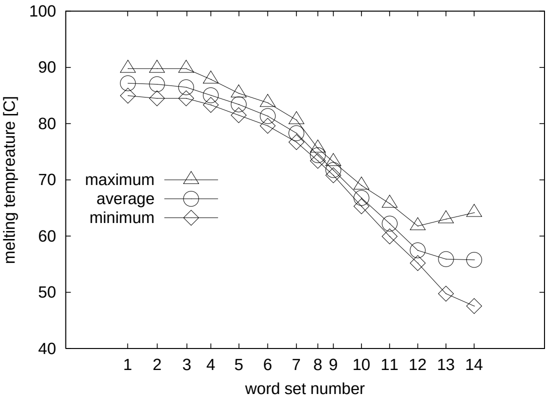

Fig. 4. Range of melting temperatures for the optimized word set at 1 M salt and 5 μM.

This behavior shows that a DNA word set optimized for molecular information processing has unique physical properties. The binding energy, melting temperature, and the nucleotide composition of the sequences are well correlated and can be varied within a small range. The optimal melting temperature (about 75 °C) corresponds to a G+C content of about 67 % for a set of 24 words of 16 mers. Variation of these values is possible to a certain degree but will be accompanied by a lesser discrimination between correct and incorrect binding processes and thus a lower fidelity of molecular computations.

There are a number of open questions concerning the potential of DNA sequences for the solution of combinatorial problems, *e.g.*, the scale-up to larger word sets, the choice of the optimal word length, a statistical analysis of the optimized sequences to study the correlation in the base composition and to study their Kolmogorov complexity. Numerical tests have shown that larger word sets (e.g. a 64 bit set) can be computed without significant reduction of the quality, whereas a restriction to a three nucleotide alphabet $\{C, A, T\}$ is accompanied by a 40% decrease of the gap δF (results not shown). The word sets designed according to this work show a preferable behavior in ongoing experimental tests, and the algorithm has successfully been applied for the design of capture probes on DNA chips, molecular beacons and primers for an isothermal DNA amplification reaction (unpublished observations).

Acknowledgements

We are grateful to J. S. McCaskill (Sant Augustin, Germany) for helpful discussions, Ivo L. Hofacker (Vienna, Austria) for his technical support with the Vienna RNA folding package, and John SantaLucia Jr. (Detroit, USA) for the permission to use his DNA energy parameters.

[1] L. M. Adleman, Science **266**, 1021 (1994).

[2] R. J. Lipton, Science **268**, 542 (1995).

[3] F. Guarnieri, M. Fliss, and C. Bancroft, Science **273**, 220 (1996).

[4] Q. Ouyang, P. D. Kaplan, S. Liu, and A. Libchaber, Science **278**, 446 (1997).

[5] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L. M. Smith, and R. M. Corn, Nucleic Acids Research **25**, 4748 (1997).

[6] D. Faulhammer, A. R. Cukras, R. J. Lipton, and L. F. Landweber, Proc. Nat. Acad. Sci. USA **97**, 1385 (2000).

[7] Q. Liu, L. Wang, A. G. Frutos, A. E. Condon, R. M. Corn, and L. M. Smith, Nature London **403**, 175 (2000).

[8] A. Marathe, A. E. Condon, and R. M. Corn, DIMACS Series in Discrete Mathematics and Theoretical Computer Science **54**, 75 (2000).

[9] R. S. Braich, N. Chevlyapov, C. Johnson, P. W. Rothemund, and L. M. Adleman, Science **296**, 499 (2002).

[10] R. Deaton, M. Garzon, R. C. Murphy, J. A. Rose, D. Franceschetti, and S. E. Stevens Jr., Phys. Rev. Lett. **80**, 417 (1998).

[11] J. SantaLucia Jr., Proc. Nat. Acad. Sci. USA **95**, 1460 (1998).

[12] J. R. Rose and R. J. Deaton,Lecture Notes in Computer Science **2054**, 231 (2001).

[13] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, Monatsh. Chem. **125**, 167 (1994).

[14] C. Cantor and P. Schimmel, Biophysical Chemistry, W. H. Freeman and Company, New York 1998, Vol. Part III, Chapt. 23.

[15] J. Ackermann, N. Loew, T. Rücker, C. Uschkereit, and F.-U. Gast, to be published.