Hamming-Chromatography

Andreas Schwienhorst[1,2], Andreas Schober[1,2], Rolf Günther[2,5] and Peter F. Stadler[3,4]

[1]Institute for Molecular Biotechnology, Jena

[2]Max Planck Institute for Biophysical Chemistry, Göttingen

[3]Institut für Theoretische Chemie, Universität Wien

[4]Santa Fe Insitute, Santa Fe

[5]Evotec BioSystems GmbH, Hamburg

Introduction:

Synthesis of random libraries of polynucleotide sequences has become a standard procedure in biotechnology; in combination with amplification methods such as the polymerase chain reaction (PCR) [1] this allows for selection experiments in which large pools of sequences are screened for a desired function. Provided the desired function can be used to design a selection constraint, these basic techniques can be combined in elegant in vitro experiments consisting of alternate cycles of selection and amplification of the selected fraction (see [2] for review).

When designing an in vitro evolution experiment a certain knowledge on the relation of the sequences and the desired function will clearly be helpful in order to decide about the minimum size of the starting pool, the optimal accuracy of the amplification steps and the strength of selection pressure in the selection steps of the procedure, or the number of iterations. In the following we will use the term fitness for the numerical value of the property that is selected for. This might be a binding constant to a prescribed substrate, the replication rate of an RNA under the action of a polymerase activity or the catalytic activity and/or the substrate qualities for a cleavage or ligation reaction.

It is useful to conceptually arrange polynucleotide sequences in such a way, that they become neighbors in an abstract "sequence space" if they differ only by a single mutation. The mapping from the polynucleotide sequences to their fitness values can then be viewed as a landscape over sequence space. Extensive theoretical studies on the form of fitness landscapes, in particular with models based on RNA secondary structures [3] have revealed that the autocorrelation function defined by

$$\rho(d) = 1 - \frac{\langle (f(x) - f(y))^2 \rangle_{d(x,y)=d}}{\langle (f(x) - f(y))^2 \rangle_{rnd}}$$

is a crucial measure for the structure of a landscape. Here $\langle (f(x) - f(y))^2 \rangle_{d(x,y)=d}$ is the average squared difference of the fitness values of sequences with mutual Hamming distance d; the denominator is the average squared distance of a random sample of sequences. The information contained in $\rho(d)$ can in most cases be reduced to the correlation length, which itself is closely related to the number of local fitness optima and the optimal mutation rate in an evolution experiment on the landscape of concern [4].

While the usefulness of information about a fitness landscape for practical applications is obvious, detailed experimental exploration of a landscape is a tedious task involving either extensive sequencing of clones obtained from a random library or extensive synthesis of a large number of distinct DNA sequences in parallel [5-7].

It is clear that the information provided by the correlation length of a landscape is of practical use only if it can be obtained easily for a given experimental system - there is no point in extensive studies of individual molecules just in order to estimate a suitable schedule for a selection experiment. Probably trying a number of them directly would be more efficient. In general testing the function , i.e. estimating the fitness of a particular sequence is easy compared to the sequencing or synthesis work required to determine which sequence is considered.

The correlation structure of the landscape can be inferred, however, without any knowledge of the sequences themselves - it suffices to determine the distances between sequences. At first glance this task does not seem to be easier. Taking into account, however, the large number of different sequences in a test tube experiment, it suffices to measure $\langle (f(x) - f(y))^2 \rangle_{d(x,y)=d}$ centered at a fairly small number of test sequences x. A correlation function can therefore be efficiently measured if one has a means of separating a pool of (rapidly synthesizable) random sequences according to their Hamming distance from a moderately small number of test sequences, thus statistically probing sequence space locally. Here we describe a chromatographic technique that is a suitable tool for this: Hamming-Chromatography.

<u>Methods:</u>

DNA molecules. DNA oligomers (Fig. 1) were synthesized on a Pharmacia Gene Assembler Plus. Oligonucleotide "T7-top" (TAA TAC GAC TCA CTA TAG) was aminoalkylated at its 5´-end using standard protocols [8].

Buffers for chromatography. Tetramethylammonium chloride (TMACl, Aldrich) was dissolved in distilled water to give an approx. 5 M solution, Norit A was added, the solution stirred and than filtered. After measuring the refractive index to determine the precise concentration [9], distilled water, sodium phosphate and SDS were added to give the desired concentrations of the 2X Hamming buffer (3 M TMACl, 0.02 M Sodium phosphate, pH 6.8 and 0.2 % SDS). The gradient was generated by mixing gradually 1X Hamming buffer with 1X Hamming buffer containing 50 % formamide.

Preparation of the affinity column. DNA oligomer "T7-top" (370 μg) previously aminoalkylated at its 5´-end was incubated with 200 μl Affi-Gel (BIORAD) in 0.1 M MOPS, pH 7.5 for 4 hours at 4 ^{o}C. After sedimentation, the Affi-Gel was incubated overnight with 0.1 M Glycinethylester at 4 ^{o}C. The Affi-Gel was washed extensively with 0.1 M and 0.001 M MOPS, pH 7.5 and packed into a small chromatography column.

Hybridization of a DNA oligomer-pool to the prepacked affinity column. 0.02 μg each of DNA oligomers T7-4, T7-21, T7-22, T7-1 and T7-0 (Fig. 1) previously labeled with ^{32}P-γ-ATP and T4 kinase [10] to a final activity of 1.5 x 10^6 cpm/μg were applied to the prepacked affinity column by circular flow-through for 30 min. at 37 ^{o}C. At the end of this annealing procedure basically no radioactivity could be detected in the flow-through.

Eluation of DNA oligomers in an "effective temperature-gradient". DNA Oligomers in principle could be eluated in a true temperature gradient. However, hence a constant temperature can be achieved much more easily and less costly than an acurate temperature gradient, we chose a formamide gradient as an "effective temperature" gradient instead. All buffer reservoires together with the chromatography column and all tubings were equilibrated to 37°C in an incubation chamber. The column was embedded in an aluminium block (10 cm X 10 cm X 10 cm) which was calculated to have a heat capacity that should maintain its temperature with a deviation of less than 0.1 °C.
DNA Oligomers were eluted in a gradient of 0 -50 % formamide in Hamming buffer. Fractions of 100 μl were collected and 50 μl-aliquots were measured in a szintillation counter.

Cloning and Sequencing of ssDNA.

Fractions from the different peaks were pooled and purified using NAP-5 columns. Eluting oligonucleotides were concentrated using a SpeedVac and sequenced by a modified version [11, 12] of the Okayama/Berg-method [13].

Results:

The correlation function is a useful means of characterizing a given landscape. It can be efficiently measured if one has a method of separating a pool of random sequences according to their Hamming distance from a moderately small number of test sequences. We have developed a type of hybridization chromatography, where a mixture of oligos (partially) complementary to a given test sequence is hybridized to the test sequence, covalently bound to a matrix. DNA Oligomers were eluted using an "effective temperature-gradient", i.e. a formamide gradient in an elution buffer containing 1.5 M tetramethylammonium salts, known to diminish the effects of GC- versus AT-pairs on the melting temperature [9, 14-17].

As a first test case we used the sequence of the top strand of the T7-Promotor imobilized to an affinity column via a previously synthesized amino-link. A mixture of (partially) complementary strands containing one 4-error mutant, two 2-error mutants, one 1-error mutant and the exact copy of the complementary sequence was quantitatively hybridized to the column. It could be shown, that the DNA-oligomers eluted in four distinct peaks the second of which was higher and broader than the other ones, reflecting the concomittant eluation of two different DNA oligomers as could have been expected from the two comigrating 2-error mutants T7-21 and T7-22 (Fig. 2). Sequencing of cloned molecules derived from pooled fractions of either of the four peaks proved to reveal the corresponding sequences.

In contrast to the artificial mixture used in the initial experiment with the T7 promotor variants, realistic mixtures originating from chemical coupling reactions of 4-nucleotide mixtures of synthons do not consist of equimolar mixtures of variants. However, hamming chromatography is most effective if the total amount of nucleic acid falling into the d-error class (measured from the reference sequence) is approximately the same for all classes. While it is technically not feasible to achieve this goal exactly, a biased synthesis with an enhanced fraction of the oligonucleotides present in the desired error-classes can approximate the desired result. Let q denote the fraction of the exact complement-nucleotide in the random pool (we assume the same fraction for all positions). Then the sequences are distributed over the error-classes according to the following Binomial distribution

$$p(d) = \binom{n}{d} q^{n-d} (1-q)^{d}$$

where n is the number of randomized positions. A good choice of q is obtained by maximizing $\sum_{d=d_1}^{d_2} p(d)$.

In general one has $d_1 = 1$ such as in the example discussed in Fig. 3. Usually one will be interested in local features of the landscape, so the upper limits d_2 will be much smaller than the number n of randomized nucleotides.

Discussion:

We have shown that the separation of a random pool of nucleic acid sequences into error classes with respect to an arbitrary reference sequence is feasible. In a model experiment a mixture of five sequences representing variants of the T7 promotor sequence was separated into four different error classes. High concentrations of tetramethyl ammonium salts were used in order to diminish the dependence of melting temperature on the GC-content presumably by binding to and thus stabilizing A:T base pairs [10, 18] although recently a sequence context dependence of melting temperatures [19] and the dependence of melting temperatures on the kind of mismatch [20] have been reported for some cases. The optimal tetramethyl ammonium concentration is dependent on the length of the oligonucleotide. For oligonucleotides on a solid support it has been suggested that the shorter the oligonucleotides the higher a concentration of tetramethyl ammonium is needed to equalize the contributions of A:T and G:C base pairs.

Shorter oligonucleotides should give better discrimination, since the relative decrease in hybrid stability is greater than for longer probes. Thus, in principle the method can only be applied to sequences with the length of some 20 nucleotides assuming that 1 % mismatch roughly reduces the metling temperature by between 0.5 and 1.4 °C [21,22,23]. This restriction can be overcome by combining blocks of sequence after chromatographing them separately in advance. Furthermore sequences to be separated should have a reasonable amount, say 50%, of homology to the reference sequence in order to hybridize at all. This does not seriously limit the applications because one is interested in the close vicinity of the reference sequence anyways.

A formamide gradient rather than a true temperature gradient although very limitted in the range of melting temperatures covered was used because of two main advantages: (i) enhanced specificity [24] and (ii) no need for special equipment besides standard chromatography technology. However, if neccessary future experiments could be performed using a combination of formamide, salt and temperature gradients.

Single-stranded DNA molecules were cloned and sequenced. Sequence analysis revealed, that oligonucleotides eluating from the column in four distinct peaks indeed fall into the different error classes as has been predicted, with the variant containing many mismatches eluating firstly and the exact complement eluating at last. Hence, hamming chromatography seem to be a technology idealy suited for experimentally characterizing landscapes of biopolymers. It possibly also could provide a useful alternative to site directed mutagenesis. Within a single synthesis <u>all</u> possible d-error mutants can be produced in parallel. Separate cloning of the error classes into an appropriate expression system, preferentially a system that allows for selection of active mutants, will allow for immediate recognition of acitve (or inactive) mutants that can be sequenced. The need for sequencing distinct mutants appears to be no disadvantage to the classical site-directed mutagenesis protocols because there sequencing has to be done as a control anyways.

Lastly the method presented here offers a way of designing sequence pools with a bias towards desired mutant spectra, thus could be useful in combinatorial library design.

Acknowledgements:

This research was supported in part by grant Sch 578/1-1 from the Deutsche Forschungsgemeinschaft (DFG) and a grant from the monastery of Andechs. The authors also would like to thank the Santa Fe Institute for providing a very stimulating atmosphere for working out part of the method described here.

References:

[1] Mullis, K.B., Faloona, S. Scharf, R., Saiki, G. Horn and H. Ehrlich (1986) Cold Spring Harb. Symp. Quant. Biol. 51: 263-273.

[2] Eigen, M. (1987) Cold Spring Harb. Symp. Quant. Biol. 52: 307-320; Kauffman, S.A. (1992) J. Theor. Biol. 157: 1-7; Szostak, J.W. (1992) TIBS 17: 89-94; Waldrop , M.M. (1992) Science 256: 1396-1397; Joyce, G.F. (1993) Currrent Biology 3:703-704; Cech, T.R. (1993) Nature 365: 204-205.

[3] Fontana, W., Konings, D.A.M., Stadler, P.F. and Schuster, P. (1993) Biopolymers 33: 1389-1404.

[4] Schuster, P., Stadler, P.F. (1994) Computers and Chemistry 18: 295-324

[5] Fodor, S.P.A., J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, D. Solas (1991) Science 251: 767-771.

[6] Southern, E.M., Maskos, U. and J.K. Elder (1992) Genomics 13: 1008-1017.

[7] Schober, A., A. Schwienhorst, A., J.M. Köhler, M. Fuchs, R. Günther, M. Thürk (1994) in: H.Reichl and A. Heuberger (Eds.) Microsystem Technologies ´94, VDE-Verlag Berlin-Offenbach, p. 381-390

[8] Schriek, U. (1989) Pharmacia LKB.

[9] Wood, W.I., J. Gitschier, L. A. Lasky and R. M. Lawn (1985) Proc. Natl. Acad. Sci. USA 82: 1585-1588.

[10] Sambrock, J., Fritsch, E.F., Maniatis, T. (1989) Molecular Cloning, CSHL Press, Cold Spring Harbor.

[11] Temsamani, J., Kubert, M., Agrawal, S. (1995) Nucl. Acids Res. 23: 1271-1272.

[12] Temsamani, J., Kubert, M., Agrawal, S. (1995) Nucl. Acids Res. 23: 1841-1844.

[13] Okayama, H., Berg, P. (1992) Mol. Cell. Biol. 2: 161-170.

[14] Shapiro, J.T., B.S. Stannard, G. Felsenfeld (1969) Biochemistry 8: 3233-3241.

[15] Orosz, J. M. and J.G. Wetmur (1977) Biopolymers 16: 1183-1199.

[16] Jacobs, K.A., Rudersdorf, R. , Neill, S.D., Dougherty, J.P., Brown, E.L., Fritsch, E.F. (1988) Nucl. Acids Res. 16: 4637-4650.

[17] Maskos, U., Southern, E.M. (1992) Nucleic Acids Res. 20: 1675-1678.

[18] Melchior, W.B. Jr. and P.H. von Hippel (1973) Proc. Natl. Acad. Sci. USA 70: 298-302.

[19] Riccelli, P.V. and A.S. Benight (1993) Nucleic Acids Res. 21: 3785-3788.

[20] Maskos, U. and Southern, E.M. (1993) Nucleic Acids Research 21: 4663-4669.

[21] Wallace, R.B., J. Shaff, R.F. Murphy, J. Bonner, T. Hirose and K. Itakura (1979) Nucleic Acids Res. 6: 3543-3557.

[22] Drmanac, R., Z. S. Trezoska, I. Labat, S. Drmanac and R. Crkvenjakov (1990) DNA and Cell Biol. 9: 527-534.

[23] Anderson, M.L.M. and B. D. Young (1991) in: B.D. Hames and S.J. Higgins (Eds.) Nucleic acid hybridisation, IRL Press, Oxford-Washington DC, p. 73-111.

[24] McConaughy, B.L., C.D. Laird and B.J. McCarthy (1969) Biochemistry 8: 3289-3295.

"T7-top" 3´– GATATCACTCAGCATAAT – 5´– NH_2

"T7-0" 5´– CTATAGTGAGTCGTATTA – 3´
"T7-1" 5´– CTATAGTGAGGCGTATTA – 3´
"T7-21" 5´– CTATAGTGAGGAGTATTA – 3´
"T7-22" 5´– CTATAGAGAGTAGTATTA – 3´
"T7-4" 5´– CTATAGAAAGGAGTATTA – 3´

Fig. 1

DNA oligomers used in the model experiment. Aminoalkylated T7-top was covalently bound to Affy-Gel (BIORAD). T7-4, T7-21, T7-22, T7-1 and T7-0 are (partially) complementary DNA oligomers capable of hybridizing to T7-top.

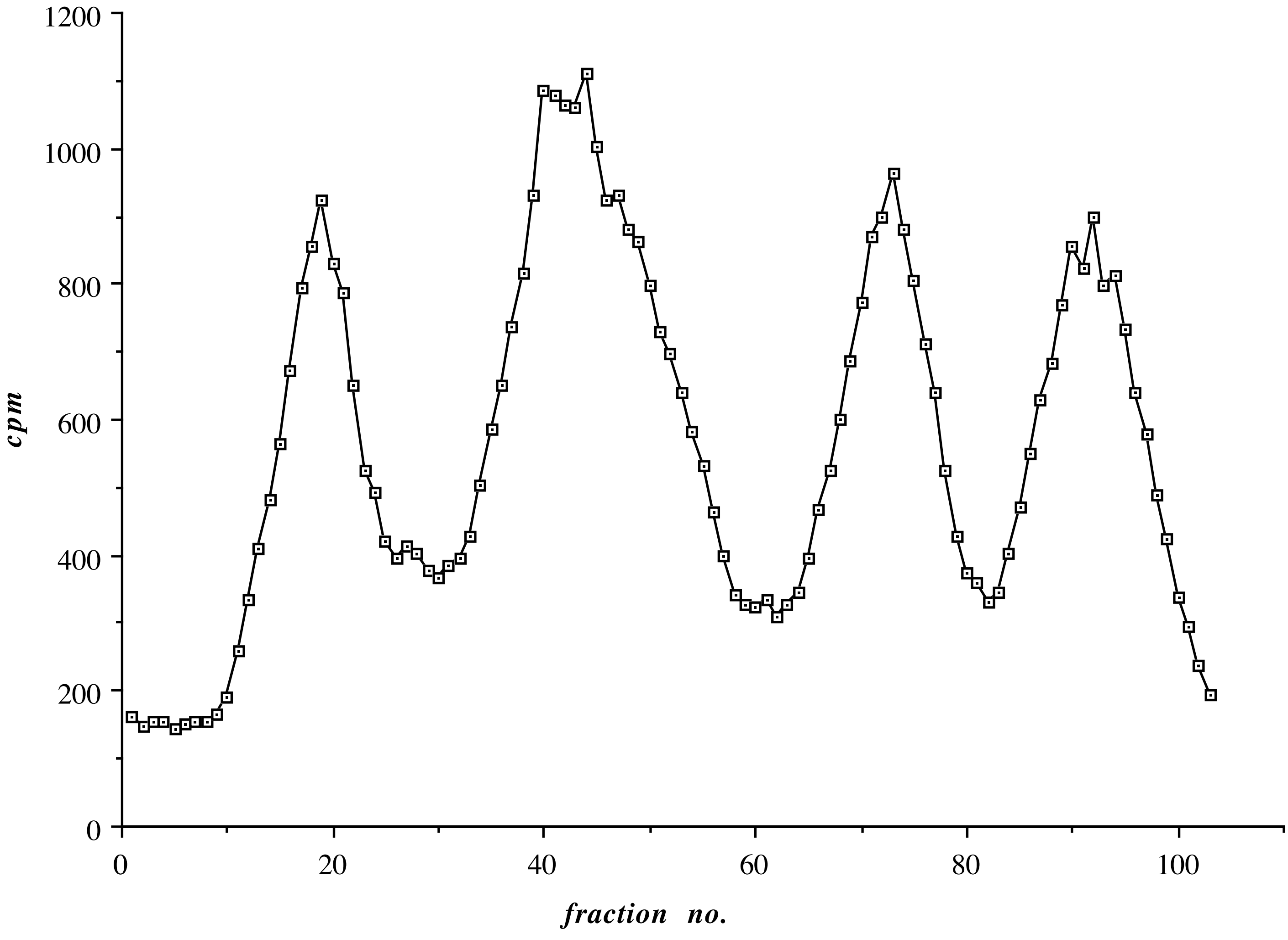

Fig. 2

Hamming chromatography of a mixture of a 4-error mutant, two 2-error mutants, one 1-error mutant and the exact copy of the complementary sequence of the T7 promotor top strand. DNA-oligomers eluted in four distinct peaks the second of which was higher and broader than the other ones, reflecting the concomittant eluation of two different DNA oligomers as could have been expected from the two comigrating 2-error mutants T7-21 and T7-22

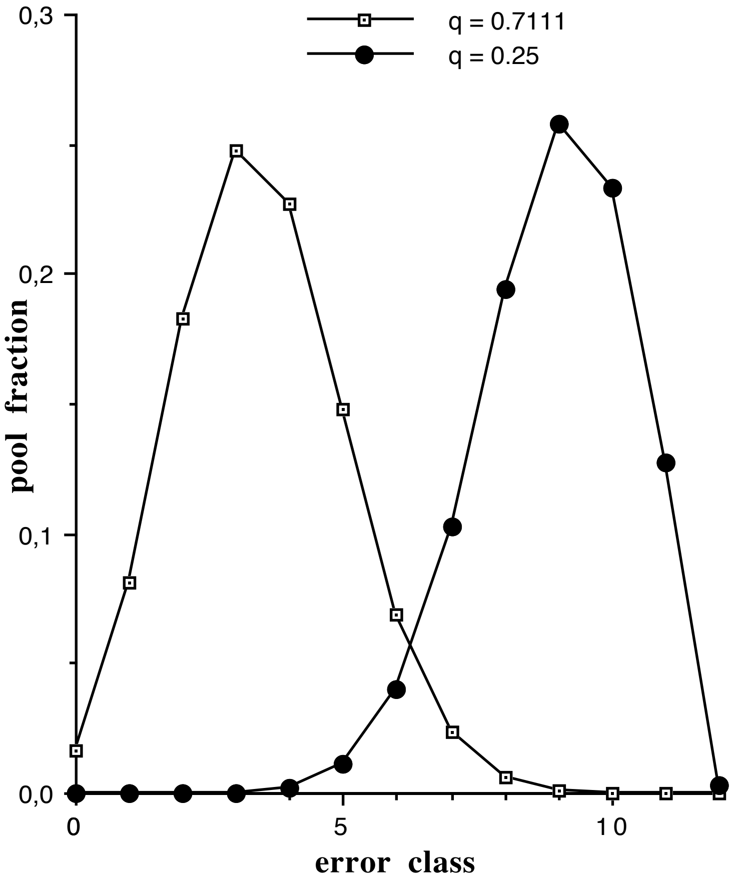

Fig. 3

Relative concentration of error classes for 12 randomized nucleotide positions. Equal fractions of the four different bases, q = 0.25 (•) yielding almost exclusively sequences with Hamming distance larger than 5, i.e. largely outside the range of interest. The best choice of q for producing predominantly the error classes 1 through 6 with approximately equal frequencies is q = 0.7111 (±).