ELSEVIER

Journal of Chromatography A, 739 (1996) 431–447

JOURNAL OF CHROMATOGRAPHY A

Fractals for multicyclic synthesis conditions of biopolymers Examples of oligonucleotide synthesis measured by high-performance capillary electrophoresis and ion-exchange high-performance liquid chromatography

Zeno Földes-Papp[a,*], Eckhard Birch-Hirschfeld[b], Holger Eickhoff[b], Gerd Baumann[c], Wei-Guo Peng[d], Thomas Biber[e], Rüdiger Seydel[e], Albrecht K. Kleinschmidt[f], Hartmut Seliger[a]

[a]*Sektion Polymere, Universität Ulm, D-89069 Ulm, Germany*
[b]*Institut für Molekulare Biotechnologie e.V., D-07745 Jena, Germany*
[c]*Abteilung Mathematische Physik, Universität Ulm, D-89069 Ulm, Germany*
[d]*Softlab GmbH, D-81677 München, Germany*
[e]*Abteilung Mathematik VI, Universität Ulm, D-89069 Ulm, Germany*
[f]*Universität Ulm, D-89069 Ulm, Germany*

Abstract

We have developed models of patterns for nucleotide chain growth. These patterns are measurable by high-performance capillary electrophoresis and ion-exchange high-performance liquid chromatography in crude products of solid-phase synthesized 30mer and 65mer oligodeoxyribonucleotide target sequences N. We introduce mathematical methods for finding characteristic values d_0 and p_0 for constant chemical modes of growth as well as d and p for non-constant chemical modes of growth (d=probability of propagation, p=probability of termination). These methods are employed by presenting the accompanying computer software developed by us in C code, Mathematica^R languages, and Fortran.

Characteristic values of the parameters d, p, and the target nucleotide length N describe the complete composition of the crude product. From this we have developed the relation $2-[N/(N-1)]/D_{\mathrm{a,\ measurable}}(N,d)$ as a universal quantitative measure for multicyclic synthesis conditions (D, fractal dimension and similarity exponent, respectively). We use this mathematical treatment to compare the efficiency of oligodeoxyribonucleotide syntheses of different target length N on polymer support materials. Further, we analyze selected syntheses of short and long oligodeoxyribonucleotides as well as single-stranded DNA sequences by well-known empirical autocorrelation, fast Fourier transformation, and embedding-dimension techniques.

Keywords: Oligonucleotide synthesis, chemical; Nucleotide sequence variability; Gene assembly; Fractal dimension; Data analysis; Polymerization processes; Enzymatic nucleotide growth; Oligodeoxyribonucleotides; Oligonucleotides; Oligoribonucleotides; DNA

1. Introduction

DNA genetic drugs are becoming the focus of the most advanced pharmaceutical research. They are used for diagnostic and therapeutic purposes, e.g. in the control of viral and neoplastic disorders [1] or in the study of biochemical functions of various gene products in vitro and in vivo [2–4]. In the field of antisense oligonucleotides fast, reliable, highly effi-

[*] Corresponding author. Present address: Department of Medical Biophysics, Karolinska Institute, S-171 77 Stockholm, Sweden.

PII S0021-9673(96)00060-X

cient analytical techniques are required for checking purity and stability of these synthesized compounds [5]. High-performance liquid chromatography (HPLC) methods are widely used to separate and purify synthetic oligonucleotide crude products up to 30mers [6]. Recent developments of quantitative interpretation of electropherograms promote separation by high-performance capillary electrophoresis (HPCE) as an analytical tool to assess the purity [7–13]. HPCE and ion-exchange HPLC can detect and quantify failed and truncated oligo-DNA sequences [5,14–18]. However, little information is so far available to interpret elution profiles in terms of characteristic synthesis parameters for small- and large-scale chemical production. We succeeded in giving a complete description of the accumulation and propagation of failure sequences during the chemical synthesis of oligonucleotides on polymer supports on the basis of fractal mathematics. This quantitation is not only of theoretical value, but has also a distinct practical importance in view of the large-scale production of oligonucleotides for diagnostic as well as therapeutic purposes [19–24]. Gram quantities up to the kilogram range [25] may be required for pharmacokinetic studies, when it comes to preclinical and clinical testing. It must be emphasized that large-scale synthesis relies on separation techniques to assess the purity of the target product. Industrial production of gene fragments and gene transcripts, which are optimally free of missense products, is without precedent in oligonucleotide chemistry [26].

The multicyclic process of oligonucleotide synthesis on solid-phase using the phosphoramidite method includes, in each cycle, detritylation, coupling and oxidation (=propagation), and capping (=termination) [27]. In this paper, we develop methods for evaluating characteristic values of d_0 and p_0 for constant, and of d and p for non-constant chemical modes of growth not related to pre-existing templates (d=probability of propagation, p= probability of termination). Furthermore, we evaluate the geometric patterns of solution profiles (speaking in terms of governing equations) and the geometric patterns of experimentally determined elution profiles by well-known empirical autocorrelation, fast Fourier transformation, as well as embedding-dimension techniques. The methods are employed by presenting the accompanying computer software in C code, Mathematica^R languages, and Fortran. Based on experiments we describe the dynamics of multicyclic growth syntheses far from equilibrium. These approaches center on fractal geometry [28]. A fractal dimension D enables comparison of multicyclic syntheses at different target length N. The mathematical relationship of universal measures for multicyclic synthesis conditions becomes constructive. The fractals can describe the defined chemical growth, and in more complexity, an equivalent of enzymatic growth as numerical, non-linear processes [29].

Experiments are provided to separate crude products by ion-exchange high-performance liquid chromatography (HPIEC) and HPCE involving data handling, for which theoretical results are included. This is practically important. From a theoretical general viewpoint we are restricted neither to target and error compounds of chemically synthesized oligodeoxyribonucleotides nor to a given time scale. Elution profiles of oligoribonucleotides synthesized in a similar multicyclic fashion and of crude products obtained during the assembly of synthetic genes are treated numerically in the same formalistic way [30].

2. Theory

In this paper we describe theoretically the computerized calculations of oligonucleotide growth in symbolic dynamics [31]. Multicyclic syntheses in an arbitrary time course are prepared on solid support and measured from elution profiles in chromatographic or electrophoretic separations. They are termed in synthesis parameters: The linear target nucleotide length is N, the probability of propagation is d generally used in the chemical literature as 'coupling efficiency', and the probability of termination is p generally used in the chemical literature as 'capping efficiency'. Several factors such as coupling efficiency during synthesis, quality of CPG-bound nucleoside, quality of reagents and nucleoside phosphoramidites, cycles and programs used in synthesis, work-up and purification protocols, etc., are influencing the content not only of the $N-1$ error sequences [34] in the crude mixture (product) of oligonucleotide preparation. d and p as well as the constant parameters d_0 and p_0 are sensitive to internal and external synthesis conditions. Therefore, optimal or standard conditions of synthesis and preparation of oligonucleotides have to be used for data evaluation. All chemical reaction steps of solid-phase oligonucleotide synthesis (e.g. detritylation, coupling, capping, oxidation) are numerically captured by proper scaling (fitting) of d and p with respect to the experimentally measured elution profile of the crude product.

The reaction front of growing chains is uniformly propagated in the chemical reaction modes of oligonucleotide growth and also in solid-phase gene assembly. The desired result of syntheses is the uniform target sequence. Thus, the model functions are applicable to other multicyclic syntheses of linear macromolecules on fixed starting sites such as oligoribonucleotide synthesis, oligopeptide synthesis, and assembly of synthetic genes. Linear means here that first of all the primary structure (sequence) has been analyzed. This generalized type of growth mechanism not related to pre-existing templates differs from growth mechanisms of polymerase chain reaction and in vitro replication. We model the dynamics without knowing details about the performance, d and p obtained from plots of absorbance versus length l of elution profiles (whereby the nucleotide length is related e.g. to retention times).

We offer here several ways to determine quantitatively target and failure sequences in oligo- and polynucleotide synthesis. Calculations of d and p values as constants d_0 and p_0 or as variables d_i and p_i, and symbolic dynamics are prerequisites for a generalization to develop universal measures of fractal dimensions D with non-constant multicyclic synthesis conditions. The approach is also useful for models of other linear macromolecular syntheses. Furthermore, empirical autocorrelation, fast Fourier transformation, and the Lyapunov exponents in embedding dimensions are applicable to this study.

2.1. Prerequisites: computerized calculations of d and p values

To describe the symbolic dynamics which assume different states at time i in a system (here, i represents the number of reaction cycles of solid-phase multicyclic synthesis), let us use the following formulations [32,33] based on [30]. $\mathbb{N}$ and $\mathbb{R}$ denote the sets of positive integers and real numbers, respectively. In brief, these states are characterized by values of the stochastic variable $\mathscr{L}(i)$. Let $I = \{i \in \mathbb{R}: 1 \leq i < N\}$, then the non-empty set I is obviously bounded above and below. After each cycle i, $\mathscr{L}$ is the realization of the nucleotide length l. The sequence l of real numbers is a map $l: \mathbb{N} \rightarrow \mathbb{R}$ where the non-empty set of its members $L(\mathbb{N}) = \{l_j: j \in \mathbb{N}\}$. For the set L we suppose that $N = \sup_{l \in L} l = \max L$ and $1 = \inf_{l \in L} l = \min L$. For the analysis [32] it will be useful to define $A^*_{260\text{ nm}}$ as the measured absorbance, e.g. at 260 nm. Then, $A^{(\text{expl})}(l,N) \equiv A^{(\text{expl})}(l)$ is the experimentally determined or theoretically calculated relative area of the peak l (integrated absorbance: absorbance·mm^2) in the chromatogram/electropherogram with largest retention of peak $l = N$. We can assume that

$$A^{(\text{expl})}(N) = \varepsilon N \prod_{i \in [1,N)} d_i \tag{1}$$

where ϵ is an empirical factor and, on the right of the product sign, d_i is the propagation probability at i to get the target sequence N. We use the Eq. 1 in the following form

$$\prod_{i \in [1,N)} d_i = \frac{A^{(\text{expl})}(N)}{\varepsilon N} \tag{2}$$

Eq. 2 is approximated (reduced) by

$$\prod_{i \in [1,N)} d_i = \frac{A^{(\text{expl})}(N) \sum\limits_{l \in [1,N]} (M(l,N)\, l)}{N} \tag{3}$$

$M(l,N)$ is the probability density for all error sequences of length l and the target sequence of length N of this homeodynamic system [30].

We obtain the synthesis parameters N, d, p directly from experimentally measured elution profiles. Let us consider in general $f(l)$ as the theoretically calculated relative area $A^{(\text{expl})}(l)$ of the peak l in the chromatogram/electropherogram. And $g(l)$ is the experimentally determined relative area $A^{(\text{expl})}(l)$ of the peak l in the chromatogram/electropherogram. We can associate with $f(l)$ and $g(l)$ the calculated error (sequence variability) ratio $\kappa_j = f_j(l)/f_j(l-1)$ and the experimentally determined error (sequence variability) ratio $\overline{\kappa_j} = g_j(l)/g_j(l-1)$, where $l =$

$\{N,N-1,\ldots,N-l_{\min}+1\}$ and $j=\{1,2,\ldots,N-l_{\min}\}$. Here $l_{\min}$ indicates the smallest peak which is still resolved in the experimentally measured chromatogram/electropherogram. Thus the absolute maximum miscalculation $\Delta z_{\max}$ is

$$\Delta z_{\max}=\sum_{j=1}^{N-l_{\min}}\left|\frac{\overline{\kappa}_j-\kappa_j}{\kappa_j}\right| \tag{4}$$

The relative maximum miscalculation $\delta z_{\max}$ is

$$\delta z_{\max}=\frac{\Delta z_{\max}}{z} \tag{5}$$

where $z=\sum_{j=1}^{N-l_{\min}}\overline{\kappa}_j$

We now discuss a way of finding the best fit of theoretically calculated elution profiles to the experimentally measured elution profile under constant and non-constant growth conditions that behaves well under certain constraints, in the following sense: $\lim_{\kappa_j\to\overline{\kappa}_j}\delta z_{\max}=0,\ \forall j\in[1,N-l_{\min}]$.

Discrepancies of many different kinds between the parameter estimation of the model and the data can be detected by studying $\delta z_{\max}$. These residuals are the quantities remaining after removal of the systematic (optimal and experimentally standardized) contributions associated with the model. When assumptions concerning the adequacy of the model are true, we expect to find that $\delta z_{\max}$ varies randomly.

2.1.1. Constant values d_0 and p_0

The model function Eq. 6a–c describes the relationships between the generated discrete sequences l for constant driven growth to produce a uniform target sequence N on solid-phase (fixed starting sites) [30]

$$\begin{aligned}M(l,N)=\left(\frac{1}{d_0}\right)^{-(l-1)}\Bigg\{&\left(\frac{1}{(1-d_0)p_0}\right)^{-1}\\&+\left(\frac{1}{p_0}\right)^{-1}\sum_{k=1}^{N-(l+1)}\left(\binom{N-(k+1)}{l-1}\right.\\&\left.\times\left(\frac{1}{1-d_0}\right)^{-(N-(l-1+k))}\left(\frac{1}{1-p_0}\right)^{-(N-(l+k))}\right)\\&+\binom{N-1}{l-1}\left(\frac{1}{(1-d_0)(1-p_0)}\right)^{-(N-l)}\Bigg\}\end{aligned} \tag{6a}$$

$$\begin{aligned}M(N-1,N)=\left(\frac{1}{d_0}\right)^{-(N-2)}\Bigg\{&\left(\frac{1}{(1-d_0)p_0}\right)^{-1}\\&+\binom{N-1}{N-2}\left(\frac{1}{(1-d_0)(1-p_0)}\right)^{-1}\Bigg\}\end{aligned} \tag{6b}$$

$$M(N,N)=\left(\frac{1}{d_0}\right)^{-(N-1)} \tag{6c}$$

where N is the number of nucleotides of a target sequence, d_0 the constant (average) propagation probability, and p_0 the constant (average) termination probability.

We verified that this relation provides accurate means to get values of d_0 and p_0 from measured elution profiles in the chromatographic separation of crude products of short target oligodeoxyribonucleotides (e.g. 30mers); N and $N-1$ peaks of the chromatogram/electropherogram are most important for the calculation of d_0 and p_0. The data evaluation was performed by repeated visual inspection of theoretical curves in comparison with the experimental elution profile [32].

Let us now introduce a more efficient method of finding the characteristic values of d_0 and p_0 from experimental elution profiles of crude products of short oligonucleotides. Eq. 3 is rewritten for constant growth ($d_i=d_0=\text{const}$)

$$d_0^{N-1}=\frac{A^{(\text{expl})}(N)\sum_{l=1}^{N}(M(l,N)\,l)}{N} \tag{7}$$

Using Eqs. 6b and 6c we write for the ratio of $A^{(\text{expl})}(N)$ to $A^{(\text{expl})}(N-1)$

$$\frac{A^{(\text{expl})}(N)}{A^{(\text{expl})}(N-1)}=\frac{Nd_0}{(N-1)(1-d_0)(p_0+(N-1)(1-p_0))} \tag{8}$$

From Eq. 8 we get

$$p_0=\frac{1}{\frac{1}{d_0}-1}\,\frac{N}{(2-N)(N-1)\kappa_1}+\frac{N-1}{N-2} \tag{9}$$

where

$$\kappa_1 = \frac{A^{(\mathrm{expl})}(N)}{A^{(\mathrm{expl})}(N-1)} \tag{10}$$

Eq. 10 is the criterion which is experimentally accessible for the data fitting procedure. For short target oligonucleotides N Eq. 10 provides adequate response to account for elution profiles obtained by HPCE and HPIEC. In modeling, the fitting procedure is an iteration between Eqs. 7 and 9 starting with an initial d_0 value. The iteration is done until the parameters d_0 and p_0 do not change any more within a chosen numerical tolerance θ.

2.1.2. Non-constant values d and p

When the target nucleotide length N is increased we supposed an exponential decrease of d_i and/or p_i [30,33]: $d_i = \alpha \exp\{-\beta(i-1)\}$, $p_i = \gamma \exp\{-\delta(i-1)\}$. This kind of function is consistent with experimental data we have so far. We show here that the target nucleotide length N can significantly affect the behavior of the experimentally measured elution profiles.

The results of this part use the recursive, non-linear model function of [32]

$$M(l,N) = M(l,N-1) + [M^{\bullet}(l-1,N-1) - M^{\bullet}(l,N-1)]d_{N-1}. \tag{11}$$

$M^{\bullet}$ is the probability density of non-terminated (error) sequences

$$M^{\bullet}(l,N) = M^{\bullet}(l,N-1)[1 - d_{N-1} - p_{N-1} + d_{N-1}p_{N-1}] + M^{\bullet}(l-1,N-1)d_{N-1} \tag{12}$$

with the initial conditions $M(1,1) = 1$, $M^{\bullet}(1,1) = 1$, and for $l > N$, $M(l,N) = 0$, $M^{\bullet}(l,N) = 0$. Here d_i and p_i are not restricted to any kind of functions within the interval $0 \leq d,p < 1$. This is of great advantage and makes the model flexible. Now we want to point out some aspects which are characteristic for this behavior and useful for data evaluation.

Here we suggest the construction of characteristic values of d_i and p_i which in turn is based on Eq. 3

$$\alpha^{N-1}\prod_{i=1}^{N-1}\exp\{-\beta(i-1)\} = \frac{A^{(\mathrm{expl})}(N)\sum_{l=1}^{N}(M(l,N)\,l)}{N} \tag{13}$$

We write for the ratios of $A^{(\mathrm{expl})}(l)$ to $A^{(\mathrm{expl})}(l-1)$

$$\frac{A^{(\mathrm{expl})}(l)}{A^{(\mathrm{expl})}(l-1)} = \frac{l\{M(l,N-1) + [M^{\bullet}(l-1,N-1) - M^{\bullet}(l,N-1)]d_{N-1}\}}{(l-1)\{M(l-1,N-1) + [M^{\bullet}(l-2,N-1) - M^{\bullet}(l-1,N-1)]d_{N-1}\}} \tag{14}$$

The data evaluation is performed as follows: In the first step constant parameters $d_0 = \alpha$ and $p_0 = \gamma$ are fitted to the chromatogram/electropherogram, i.e. to the ratio of $A^{(\mathrm{expl})}(N)$ to $A^{(\mathrm{expl})}(N-1)$ (see Section 2.1.1). The second step is to find proper exponents β and δ consecutively fitted to the ratios of peaks with $l = \{N, N-1, \ldots, N - l_{\min}\}$ for these initial values of α and γ. For this purpose iteration is done between Eqs. 13 and 14 controlled by visual inspection of theoretical elution profiles. In the following steps, the second step is repeated with increasing and/or decreasing values of α. Then, the best fit resulting is checked again by increasing and/or decreasing values of γ, followed in each case by the operations of the second step. However, this procedure is time consuming as it does not provide explicit formulations for the measured ratios $A^{(\mathrm{expl})}(l)$ to $A^{(\mathrm{expl})}(l-1)$.

2.2. The universal fractal measures D under non-constant multicyclic synthesis conditions

Beyond the description of chemical solid-phase oligonucleotide preparations we are interested in the more general application of fractal concepts [28,35,36] to multicyclic syntheses. This part of the paper is intended to build up the theoretical framework necessary to quantitatively analyze the complex behavior of multicyclic syntheses, as they occur in biology, in terms of a generalized dimension D. By the theoretical results presented here and based on the work in [30] and [32] we gain access to experiments.

In the biochemical experiment of a multicyclic synthesis (e.g. oligonucleotide synthesis) the experimental result of interest can be the relative yield of product. This would typically be calculated from the relative area of peaks $A^{(\mathrm{expl})}(l,N)$ in the chromatogram/electropherogram (integrated absorbance: absorbance·mm^2). We consider here the question: Is there evidence for any real difference among the relative target yield values (peak N in the chromatogram/electropherogram) associated with different target length N of a multicyclic synthesis? Essentially the fractal analysis of multicyclic synthesis determines whether the discrepancies between relative target yields (relative areas of target peak N) are greater (or smaller) than could reasonably be expected from the variation that occurs within the influence of target length N on the probability distribution.

The values of fractal dimension $D_{\mathrm{a,\ measurable}}(N,d)$ for a multicyclic synthesis can be computed [30] from

$$D_{\mathrm{a,\ measurable}}(N,d) = 2 - N\left(\frac{\partial(\ln\{1 - M(N,N)\})}{\partial N}\right) \tag{15}$$

This relationship is a general synthesis expression of $D_{\mathrm{a}}(N)$; the multicyclic synthesis can be carried out on fixed starting sites or in solution [32]. The formula (Eq. 15) shows that we handle the data in the following manner: Each individual result $M(N,N)$ is regarded as being made up of a point in a log–log plot: $1 - M(N,N)$ versus N. Now, if we imagine the target length N taken very large, the discrete steps associated with the curve become very small. In the limit, as the step becomes infinitesimally short, we can verify the set of points to be represented by a continuous curve. The fractal dimension we discuss measures numerical properties of such sets of points [30,32]. Multicyclic syntheses are naturally expressed by the exponent $D_{\mathrm{a,\ measurable}}(N,d)$; they are comparable in terms of their dynamics $D_{\mathrm{a,\ measurable}}(N,d)$ [32,33].

Supposing a constant growth by the propagation probability $d_i = d_0 = \mathrm{const}$ we get from Eq. 15 the practical relationship of the dynamics for each target length N [32]

$$D_{\mathrm{a,\ measurable}}(N,d_0) = 2 - N\frac{d_0^{N-1}}{1 - d_0^{N-1}}\ln\frac{1}{d_0} \tag{16}$$

To understand what is being proposed in the experiments shown in Section 4, it is helpful to have in mind the following consequences of Eq. 16. Let the yields of target sequences N generated by a constant multicyclic growth be expressed as 50% of the maximum that is theoretically attainable. Hence, we get from Eq. 16

$$D_{\mathrm{a,\ measurable}}(N,d_0) = \frac{N^2(2 - \ln 2) - N\ln 2 - 2}{N^2 - 1} \tag{17}$$

Obviously, the values of D_{a}(N) are associated with a non-linear relation to the target length N (Eq. 17). Although the yields of target sequences are equal, the dynamics of the multicyclic syntheses expressed in terms of $D_{\mathrm{a}}(N)$ are different. This phenomenon reflects the fact that yields or values of d as defined here in a multicyclic synthesis, do not allow quantitative characterization of the dynamics.

Here we extend $D_{\mathrm{a.\ measurable}}(N,d)$ to a non-constant growth. We found by experiments (see Section 4) that chemical oligodeoxyribonucleotide synthesis can be effected by an exponential propagation probability function $d_i = \alpha\exp\{-\beta(i-1)\}$. In just the same way in which Eq. 16 is obtained from Eq. 15, Eq. 18 is obtained from Eq. 15 with $d_i = \alpha\exp\{-\beta(i-1)\}$

$$D_{\mathrm{a,\ measurable}}(N,d) = 2 - N\left(\frac{\dfrac{-\partial\left(\prod_{i=1}^{N-1}\alpha\exp\{-\beta(i-1)\}\right)}{\partial N}}{1 - \prod_{i=1}^{N-1}\alpha\exp\{-\beta(i-1)\}}\right) \tag{18}$$

where $d = d_i$
Then,

$$D_{\mathrm{a,\ measurable}}(N,d) = 2 - N\left(\frac{\alpha^{N-1}(-3\beta + 2\beta N - 2\ln\alpha)\exp\left\{\dfrac{-\beta(1-N)(2-N)}{2}\right\}}{2\left(1 - \alpha^{N-1}\exp\left\{\dfrac{-\beta(1-N)(2-N)}{2}\right\}\right)}\right) \tag{19}$$

Eq. 19 is a model for data from non-constant multicyclic synthesis conditions. If we choose $\beta \to 0$ in Eq. 19, we get the fractal dimension for constant growth given by Eq. 16.

Since none of the models (Eqs. 16, 17, 19) appears independent of N, special care must be taken to eliminate the influence of N when the influence of d on multicyclic syntheses is described at different target length N. Here, in comparison of constant with non-constant growth, the method of analysis is as follows [32]. From Eq. 19 we compute an idealized multicyclic synthesis without error production, i.e. with synthesis of the target only, for $\beta \rightarrow 0$ and $\alpha \rightarrow 1$. The dynamics of such a synthesis only depends on N

$$D(N) = 2 - \frac{N}{N-1} \tag{20}$$

Hence, the quantity

$$D_a(d) = \frac{2 - \dfrac{N}{N-1}}{D_{a,\ \mathrm{measurable}}} \tag{21}$$

is characteristic for the influence of d on the dynamics of a multicyclic synthesis.

2.3. Further methods in modeling the dynamics of multicyclic synthesis on solid-phase (or in solution)

We use here well-known methods of empirical autocorrelation, fast Fourier transformation, and embedding-dimension techniques. Geometric patterns of solution profiles calculated according to our equations (see Section 2.1) and geometric patterns of experimentally determined elution profiles (see Fig. 3) are evaluated. For this purpose we take $A^{(\mathrm{expl})}(l,N)$ as an observable in the elution profile. The vector $\vec{A}^{(\mathrm{expl})}(l,N)$ stands for experimentally or theoretically verified values of relative peak areas in a given chromatogram/electropherogram with largest retention of peak $l = N$. Let us take the following description of the elution profile of a multicyclic synthesis

$$\vec{A}^{(\mathrm{expl})} := \vec{A}^{(\mathrm{expl})}(l,N) - \vec{A}^{(\mathrm{expl})}(l+1,N) = \begin{pmatrix} t_0 \\ t_1 \\ \vdots \\ t_j \\ \vdots \\ t_{N-2} \end{pmatrix} \tag{22}$$

We recognize the difference between $\vec{A}^{(\mathrm{expl})}(l,N)$ and $\vec{A}^{(\mathrm{expl})}(l+1,N)$ by considering it as a signal $f(t)$ with evolution in time t

$$f(t) := \vec{A}^{(\mathrm{expl})^{\mathrm{T}}} \tag{23}$$

where superscript T denotes the transpose of the vector. The function $f(t)$ is represented by the vector $\vec{Y}$ that has the components $y_{t_0} = y_0 := f(0), y_{t_1} = y_1 := f(\Delta t), \ldots, y_{t_{N-2}} = y_{N-2} := f((N-2)\Delta t)$. The components are obtained by sampling $N-1$ scalar measurements. The sampling intervals are $\Delta t = 1$.

2.3.1. Empirical autocorrelation of multicyclic synthesis

In optimized (standardized) multicyclic syntheses (preparations) of oligonucleotides considered here, we can assume that error sequences are not independently distributed. Therefore we use a modeling procedure that reflects this situation. The autocorrelation takes into account that groups of time-dependent observations are similarly related in probability if they are similarly spaced in time. The empirical autocorrelation coefficient ϱ_{r+1} [37] is defined here by

$$\varrho_{r+1} = \frac{\dfrac{1}{N-(1+r)} \sum_{j=0}^{N-(2+r)} (y_j - \bar{y})(y_{j+r} - \bar{y})}{\dfrac{1}{N-1} \sum_{j=0}^{N-2} (y_j - \bar{y})^2} \tag{24}$$

where $\bar{y} = (N-1)^{-1} \sum_{j=0}^{N-2} y_j$, and $r = 0,1,2,\cdots$. An autocorrelation cannot be larger than $+1$ or smaller than -1. Obviously ϱ_1 is always equal to $+1$. This makes sense because the series is perfectly correlated with its (unlagged) self. When the empirical autocorrelation coefficients for the time series data at lags $0,1,2,\ldots$ are plotted against r, the value of the lag, we get the empirical (sample) autocorrelation function. Serial correlation is mostly expected at low lags. Thus we plot ϱ_{r+1} only for $r \leqslant (N-1)/2$. The shape of the lag-r empirical autocorrelation function will help us to specify elution profiles in assessing the quality of changes. For example, the quality of solutions (speaking in terms of governing equations) can be distinguished by their geometric shape (pattern): The solution profile is stationary, regular (periodic), or irregular [38].

2.3.2. Fast Fourier transformation of multicyclic synthesis

We analyze the time behavior [38] of the signal $f(t)$ (Eq. 23) at sampling intervals $\Delta t = 1$ by

$$f(t) = f_{\nu j} := \frac{a_0}{2} + \sum_{\nu=1}^{(N-1)/2} (a_\nu \cos(\nu\omega t_j) + b_\nu \sin(\nu\omega t_j)) \qquad (25)$$

where $\omega = 2\pi/T$ with period $T = 2\pi$. In the period T there are $N-1$ functional values y_{t_j}, thus $T = (N-1)\Delta t$. The Fourier coefficients a_ν, b_ν tell us approximately how much the various frequency components ν contribute to $f(t)$. The amplitude $|c_\nu|$ is associated with $|c_\nu| = \sqrt{a_\nu^2 + b_\nu^2}$.

2.3.3. Embedding-dimension technique for multicyclic synthesis

For a multidimensional dynamic system we need simultaneous recording of the values of all variables of the system. For the homeodynamic system of multicyclic synthesis we do not know the actually required number of variables. We assume that the dynamics of this system is well described by a single variable even in a multidimensional case. Here we propose reconstructing a multidimensional description of the dynamics from the variable y_{t_j} and using the embedding-dimension technique [39].

Now, $y_{t_j} = f(t_j)$ is a variable of the homeodynamic system of multicyclic synthesis that can be easily measured, or theoretically verified, by elution profiles (see Section 2.1). The time series data of y_{t_j} generate a multidimensional embedding space. For this purpose, for some embedding dimension m the values of y_{t_j} are grouped in a consecutive fashion to form vectors $g\vec{y}_{m+j} \in \mathbb{R}^m$, where $\vec{y}_{m+j} = (y_{t_j}, y_{t_{j+1}}, \ldots, y_{t_{j+m-1}})^{\mathrm{T}}$ gives the coordinates of a single point in the reconstructed m-dimensional space. Thus, the dynamics is embedded in an m-dimensional space. The time evolution of the system is analyzed by running the procedure for successive vectors $\vec{y}_{m+j}$, where $j = 0,1,2,\ldots$. It is mathematically proven that in the embedding space the dimension $m = 2s + 1$ should be used to completely represent the dynamics of the real system with s dimensions [40]. In practice, the effective dimensionality may be smaller than that of the original dynamic system [38].

The sequence of values $\vec{y}_{m+j} \in \mathbb{R}^m$ generated is called the reconstructed reference trajectory. The behavior of trajectories in the embedding space has the same geometric and dynamic properties as the trajectories in the state space of the real system. According to [38] we use the largest Lyapunov exponent λ_1 to measure an exponential stretching of the scalar distance $|\vec{r}(t)|$ to the initial scalar distance $|\vec{r}_0|$ of the perturbed trajectory away from the reference trajectory. At each point nearby the reference trajectory the Lyapunov exponent is calculated in the m-dimensional embedding space

$$\lambda_1 = \lim_{t\to\infty} \left[\lim_{|\vec{r}_0|\to 0} \left(\frac{1}{t} \ln \frac{|\vec{r}(t)|}{|\vec{r}_0|} \right) \right] \qquad (26)$$

The Lyapunov exponents λ_1 are real numbers.

3. Experimental

3.1. Materials and syntheses

3.1.1. Reagents

5′-O-Dimethoxytrityldeoxynucleoside-3′-O-(2-cyanoethyl)-N,N-diisopropyl phosphoramidites were acquired from Applied Biosystems, MWG Biotech, Millipore, and Roth. Standard solutions of oligodeoxyribonucleotide synthesis (activation, capping, oxidation, detritylation) were also purchased from these companies. Phosphoramidites were dissolved in acetonitrile, the water content of which was ≤ 10 ppm (HPLC-grade acetonitrile from Merck) in final concentration of 0.1 *M*. Acetonitrile (impurities ≤ 30 ppm) was purchased from Roth (Germany). Other reagents were analytical-reagent grade.

Water used for sample preparation and mobile phases was purified by a Milli-Q Plus water purification system (Millipore, Bedford, MA, USA). Sample solutions were filtered through a Millipore filter (pore size 0.45 μm; Millipore).

3.1.2. Chemical solid-phase synthesis of oligodeoxyribonucleotides

Solid-phase oligodeoxyribonucleotide syntheses were carried out on automated synthesizers (Applied Biosystems Model 380B and 394 DNA synthesizers) according to 0.2-μmol cycles. In the case of polystyrene-grafted polytetrafluoroethylene support P_{29} [41], coupling and washing steps for the initial five elongations were performed during double the time (60 s) as compared with the standard synthesis cycles: small scale cyanoethyl cycle 103a of version 2.0 or 2.01 for ABI 380B synthesizer (30mer syntheses) or 0.2-μmol cyanoethyl cycle for ABI 394 synthesizer of Model 392/394T system software ROM version 2.0 (65mer syntheses). An amount of 28% aqueous ammonia solution was added to the protected oligodeoxyribonucleotides bound to the solid-phase material. The cleavage from the solid support and the deprotection was completed after incubation for 5–12 h at 55°C.

Estimation of repetitive trityl yields was done by relating the 495-nm absorbance of the solution of the individual detritylation steps to that of the detritylation of the support-bound nucleoside as described in [32]. Oligonucleotide concentrations were determined in crude products after cleavage from solid-phase [32].

The 5′ d CAA CTG ACT GGT CAA CGT CTG CGT GAA GGT 30mer heterooligonucleotide sequence synthesized is the initial part of the codogenous strand of the human γ-lipotropin gene [33]. Further, the oligomer dC_{65} was synthesized.

3.1.3. Solid-phase materials for chemical synthesis of oligodeoxyribonucleotides

Syntheses were done on the polystyrene-grafted polytetrafluoroethylene support P_{29} with nucleoside loading of 11–18 μmol nucleoside/g support. Support functionalization and application of this support for routine small-scale preparations of average-size oligonucleotides as well as of long sequences are reported in [42].

3.2. Instrumentation, chromatographic and electrophoretic conditions

Relative quantification by referring to the integrated absorbance of the peaks was carried out throughout this study. The comparison of the absolute absorbance values of corresponding peaks beyond the chromatogram/electropherogram is almost meaningless. Every effort has been made to minimize the overlap of peaks by running optimized conditions. Further, the computerized capillary electrophoresis system provides an automated method for the Gaussian deconvolution of overlapping peaks. The proper estimation of the baseline was visually controlled for all chromatograms and electropherograms. A few electropherograms had to be reintegrated.

3.2.1. Ion-exchange high-performance liquid chromatography

The experiments were performed with a Bio-Rad Model 2700 (Bio-Rad, CA, USA) with Software Series 800 HRLC-System, Version 2.30.1a. The apparatus is equipped with column oven and with the Automatic Sampling System Model AS-100T HRLC. The Bio-Rad UV–Vis detector UV-1806 was used. The anion-exchange column Mono Q HR 5/5 was purchased from Pharmacia LKB Biotechnology (Uppsala, Sweden). The following experimental conditions were found to be optimal for separation of N and $N-1$ peaks: flow-rate, 1 ml/min; temperature, 60°C; eluent, buffer A: 10 m*M* NaOH–0.1 *M* NaCl (pH 11), buffer B: 10 m*M* NaOH–1 *M* NaCl (pH 11). Linear gradients of 0% to 100% buffer B were applied over 30 and 40 min, respectively. The separation was detected by measuring the absorbance at 260 nm. Regeneration of the column was performed by regularly washing to 0% buffer B during 20 min.

3.2.2. High-performance capillary electrophoresis

Analyses were carried out on a Beckman capillary electrophoresis system, Model P/ACE 5510 (Beckman Instruments, Palo Alto, CA, USA) at 30°C equipped with a diode-array detector and software Beckman System Gold Version 8.1. The following experimental conditions were used: detection, 254 nm; injection, electrokinetic, 5 s, at a voltage of 10 kV; (−) polarity; separation at 230–320 V/cm; running buffer, 250 m*M* Tris–borate with 7 *M* urea, pH 8.4. The capillaries (type ssDNA 100, 100 μm I.D.) with linear 12% polyacrylamide gel were purchased from Beckman. The total length of the capillary was 47 cm. The window was commercially cut off with a blade at 40 cm. The capillary was conditioned and rinsed in a set procedure at the beginning and end of each measurement in order to establish reproducible and robust results. For the conditioning the capillary was filled with new buffer before each run; the buffer was changed after each run. The electrophoretic conditions are similar to those of [43], where the base numbering of FITC-$p(dT)_{16-500}$ was counted for each peak on the separation pattern resolved up to 500 bases of oligodeoxythymidic acids (FITC, fluorescein isothiocyanate).

3.3. Software

Throughout this study a software program based on the program in [33] was written to generate automatically elution profiles of a multicyclic synthesis on fixed starting sites from individual integrated absorbance of peaks measured. This program is based on the results of Eqs. 7–10 and computes characteristic values d_0 and p_0. The program was written in C code compiled with TurboC+ + from Borland, USA. The program is combined in a single, highly integrated software package MicroCal ORIGIN, Version 3.5 (Scientific and Technical Graphics in Windows), MicroCal Software, MA, USA. The input data for a simulation are the target length N, the experimentally determined values $A^{(\mathrm{expl})}(N,N)$ and $A^{(\mathrm{expl})}(N-1,N)$, the chosen numerical tolerance θ, and an initial d_0 value. The program checks whether the initial d_0 input is realistic, i.e. the calculated initial p_0 is lower than 1.0; otherwise it claims a new initial d_0 input. The results of simulation are shown on the monitor in the form of a table containing the final values of d_0 and p_0, as well as the experimentally determined and calculated values of $A^{(\mathrm{expl})}(N,N)$ and $A^{(\mathrm{expl})}(N-1,N)$, and their absolute and relative differences. For the final values of d_0 and p_0 the output file given by the user is stored and contains the probability density (relative yield) of target and error sequences according to their nucleotide length, relative yields weighted to nucleotide length, $A^{(\mathrm{expl})}(l,N)$ values, the probability density of undeleted (within the error sequence) immediately terminated error sequences and of error sequences not immediately terminated. The output data file can be loaded into the computer for graphic representation.

The use of Mathematica^{R} system (a system for doing mathematics), Enhanced Version 2.2 for Microsoft Windows (Wolfram Research, Champaign, IL, USA) is becoming increasingly popular in the mathematics and science fields. Here we apply Mathematica's programming methods to the construction of the actual programs for computing characteristic constant values of d_0 and p_0 (Eqs. 7–10) and characteristic non-constant values of d and p (Eqs. 13, 14) from experimentally determined elution profiles using Eqs. 4 and 5. The actual programs include symbolic, numerical and graphical computation. The external programs are controlled by Mathematica^{R} control language for external processes and are implemented in high-level shell for file and data manipulation. The obtained results (see Fig. 1 and Fig. 2) are examples of the powerful software tool Mathematica^{R} and its useful problem-solving techniques. The extended software package for the quantitative analysis of oligonucleotides and single-stranded DNA sequences in the crude product of chemical synthesis will be available from Wolfram Research, 100 Trade Center Drive, Champaign, IL 61820-7237, USA, as well as from its network worldwide [44].

The kernel part of the Fortran program for computing the largest Lyapunov exponent λ_1 (see Section 2.3.3) from the time-series data was taken from [45]. However, Wolf et al. [45] provided only a short torso-algorithm. The automatic choice of the parameters necessary for computation is an improvement of the program used here. Further, in time-series data containing very small elements, *di* (distance before time evolution between the point on the reference trajectory and the corresponding point on the perturbed trajectory) or *df* (distance after time evolution between the point on the reference trajectory and the corresponding point on the perturbed trajectory) may be set to zero; the program used ensures positive values of *df* or *di*. The embedding dimension dim is determined automatically. The integer variable dim is increased from 2 to 15, i.e. $2 \leq \mathrm{dim} \leq 15$. The exponent λ_1 from the time-series data is separately computed for each embedding dimension and compared with the Lyapunov exponents computed just before. The computation is finished if λ_1 of the embedding dimension dim + 1 is close to λ_1 of dim.

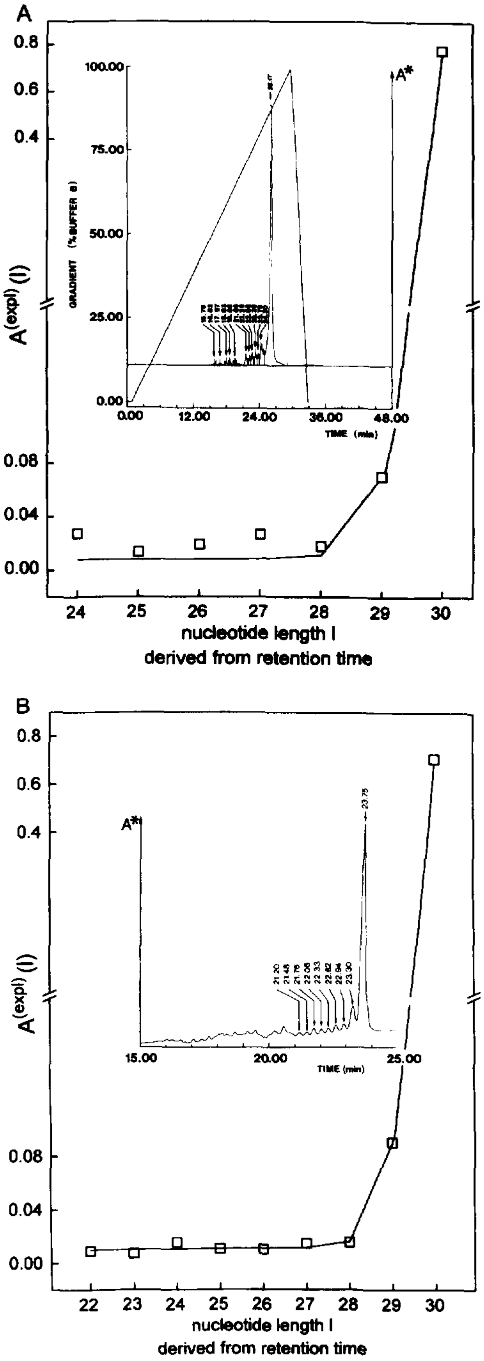

Fig. 1. Separation (insets) of the crude product of the 30mer heterooligodeoxyribonucleotide synthesis by (A) ion-exchange high-performance liquid chromatography and (B) high-performance capillary electrophoresis. The theoretical curve (———) was automatically fitted to the experimental values (□) by proper setting of synthesis parameters d_0 and p_0 for the target length $N=30$. (A) $d_0=0.986131$, $d_0=\bar{d}$ (average); $p_0=0.799990$, $p_0=\bar{p}$ (average). (B) $d_0=0.980681$, $d_0=\bar{d}$ (average); $p_0=0.796225$, $p_0=\bar{p}$ (average). See text for explanation.

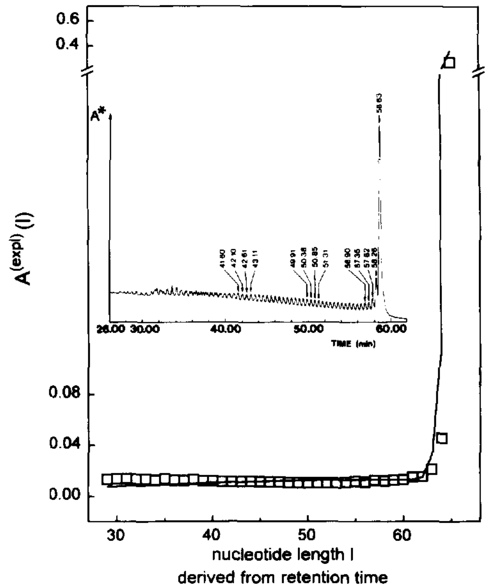

Fig. 2. Separation (inset) of the crude product of the dC_{65} synthesis by high-performance capillary electrophoresis. The experimental values (□) of the electropherogram are compared with theoretical values (———). Calculated synthesis parameters for the target length $N=65$: $d_i=0.998\exp\{-0.0006(i-1)\}$, $\bar{d}=0.97938$ (average) and $p_i=0.995\exp\{-0.00705(i-1)\}$, $\bar{p}=0.80363$ (average). See text for explanation.

Generally, the longer the time-series data are given, the better the approximation of λ_1. This program should not be used with time-series data shorter than 400 elements. For determining chaos the time-series data should contain several oscillations. For rather short time series (around 400 or less), or time series with few oscillations it is useful to compare the results of the program used with the results of other chaos-indicating techniques, e.g. with those of fast Fourier transformation or computation of the autocorrelation function.

The algorithm used for computing the fast Fourier transformation is a subroutine taken from [46]. The algorithm used for computing the autocorrelation function was also written as a subroutine in Fortran code.

4. Results and discussion

For our experiments in this pilot study we used two different values of target length N of chemically synthesized oligodeoxyribonucleotides. The experimental data for the separation of crude products are presented in Fig. 1 and Fig. 2.

The inset of Fig. 1A is the UV chromatogram of conventional HPIEC with an analytical scale column. The corresponding UV electropherogram is shown in the inset of Fig. 1B. Using the anion-exchange column Mono Q HR 5/5 from Pharmacia LKB Biotechnology the N and $N-1$ peaks for 30mer crude products are resolved within acceptable limits [32] by HPIEC (Fig. 1A, inset). Although there is a good overall agreement between the separation by HPIEC and HPCE methods, some differences do occur even at the target length $N=30$. The resolution of smaller peaks has been achieved by HPCE (Fig. 1B, inset). In all HPCE experiments it was important to condition and regularly fill the capillary with new buffer at start up, between runs, and at the end of the day to get reproducible, robust results. Under the chromatographic or electrophoretic conditions (unmodified bases, gradient conditions, mobile phase, temperature, pH, etc.) used by us, base composition [47,48] of resolved HPICE and HPCE does not influence the results in the crude products. Very recently, a lot of work has been done addressing experimentally the high-resolution analysis (and purification) of chemically synthesized oligonucleotides with strong anion-exchange HPLC [18]; very high-resolution separations were obtained by using the NucleoPac PA-100 column. In cooperation with Dionex, this column will be applied to our modeling strategy.

We modeled the measured elution profiles by our method of automatically calculated synthesis parameters d_0 and p_0 in an iterative way (see Section 2.1.1). The following parameters of Eq. 6 were obtained from the experimental data: In case of HPIEC $d_0=0.986131$ (here $d_0=\bar{d}$ (average)) and $p_0=0.799990$ (here $p_0=\bar{p}$ (average)) were calculated, whereas we found $d_0=0.980681$ (here $d_0=\bar{d}$ (average)) and $p_0=0.796225$ (here $p_0=\bar{p}$ (average)) in case of HPCE. The target length N was 30. The good correspondence between experimental (open squares) and calculated (solid line) curves (elution profiles) is evident from Fig. 1. From Eq. 8 it is apparent that the values of d_0 and p_0 depend on the ratio of relative peak areas N to $N-1$. In a recent study from another laboratory, the $N-1$ population of error sequences was analyzed by cloning and sequencing [49] for a 25mer oligodeoxyribonucleotide synthesis. No error nucleotides were found in the last four positions at the 5′-end of $N-1$ error sequences. From this it was concluded [49] that the coupling/capping/detritylation/oxidation reactions proceed with lower efficiency at the 3′-end (i.e. at the beginning of the synthesis) than at the 5′-end, proceeding at 100% efficiency at the end of the synthesis. However, this does not necessarily apply to the data presented in [49] because the chain length distribution of the non-linear growth system is not represented by one error population only, e.g. by $N-1$ error sequences. Although the complete composition may be difficult to assess by experimental means, from a theoretical point of view local nucleotide results cannot be expected to give a global picture of a non-linear growth system.

So far there is no statistically significant difference between the obtained parameters d_0 and p_0 for the two separation methods of HPIEC and HPCE under the experimental conditions used. As a consequence of the observed separation of smaller peaks by HPCE we decided to separate the crude product of the 65mer synthesis by this method.

In the inset of Fig. 2 the resolving power of HPCE is demonstrated for the crude products of a 65mer oligodeoxyribonucleotide synthesis. The UV electropherogram separated well peaks from N to $N-36$. The measured data (open squares) discredit the constant growth during multicyclic synthesis of the 65mer. The distribution of target and error sequences (solid line) can be generated by calculating d and p from all separated peaks in the electropherogram using Eqs. 13 and 14. Characteristic probability values of $d_i=0.998\exp\{-0.0006(i-1)\}$, $\bar{d}=0.97938$ (average) and $p_i=0.995\exp\{-0.00705(i-1)\}$, $\bar{p}=0.80363$ (average) were obtained by the data fitting procedure for non-constant synthesis conditions (see Section 2.1.2). The theoretical distribution of Fig. 2 can be regarded as a good approximation for the real, experimental situation of the 65mer synthesis. Thus, we conclude that further differences exist among the 30mer constant and the 65mer non-constant growth processes of Fig. 1 and Fig. 2 presented herein. Analyzing the by-products of the 30mer and 65mer syntheses, it is seen that failed and truncated sequences 'decay' to the left of the target sequence in the electropherograms. This is the first time that it is made possible to quantify automatically all individual failure sequences arising during solid-phase oligonucleotide synthesis with respect to their number and molar fraction in a sophisticated way. Clearly, the experimental methods available do not separate individual molecular species, but at least they separate the mixture into individual chain lengths. Therefore, the 'decay' of failed and truncated sequences cannot be kinetically expressed by an exponential law from measured elution profiles. The data analysis we have been developing (see Section 2) provides straightforwardly the basis for quantifying differences.

Difference between the relative areas of target peaks N [$A^{(\text{expl})}(N,N)$] of the measured 30mer and 65mer elution profiles can be detected in the chromatogram/electropherograms of Fig. 1 and Fig. 2. At $N=30$ we measured about 74.59% of the total peak area and at $N=65$ about 27.43% of the total peak area. The residual contains systematic tendencies evaluated by computer simulations of elution profiles: (i) For the same target length N the drop in $A^{(\text{expl})}(N,N)$ values is caused by decreasing d and/or p values. The influence of d is much stronger than the influence of p. However, the relative yields of the target N, i.e. the $M(N,N)$ values, depend only on d (Eqs. 6c and 11). Thus, in practice it would be unwise to take individual $A^{(\text{expl})}(l,N)$ values or their ratios as measures for the yields of target or error sequences. Nevertheless, the data vector $\vec{A}^{(\text{expl})}(l,N)$ reflects the character of the dynamics of multicyclic syntheses (see below). (ii) We have seen from the experimental data of Fig. 1 and Fig. 2 that the average $\bar{d}$ values are the same whereas the relative (and absolute) yields of error sequences are increased in Fig. 2 as compared with Fig. 1. For increasing target length N the relative target yields decrease at same $\bar{d}$ values (Eqs. 6c and 11). Thus, the propagation probability (d) called 'coupling efficiency' in the chemical literature does not tell us anything on the time course of a multicyclic synthesis. The same is true for relative (absolute) yield values. Values of d and yield values only cover the influence on any reaction step, but they neglect the portion which is independent on synthesis length N. Therefore, d values and yield values are not suitable for an accurate comparison of multicyclic syntheses at different target length N.

We have developed methods (see above) for calculating the length composition of the mixture of oligonucleotides – target sequences, truncated and failure chains – that arises during the polymer-supported synthesis of oligonucleotide sequences of given target length. The reaction front of growing chains is uniformly propagated in the chemical reaction modes of oligonucleotide growth. The desired result of syntheses is the uniform target sequence. Thus, the model functions are applicable to other multicyclic syntheses of linear macromolecules on fixed starting sites such as chemically modified phosphorothioates DNA synthesis. Linear means here that first of all the primary structure (sequence) has been analyzed. Theoretical analysis of properties of the model function for DNA growth in the mathematical sequence space [33] shows, in principle, the validity of this function as applied to branched pseudo-random walk. This generalized type of growth mechanism not related to pre-existing templates differs from growth mechanisms of polymerase chain reaction and in vitro enzymatic replication.

In addition to these observations, three other quantities are useful for the analysis of multicyclic syntheses. These are the dimensions $D_{\text{a, measurable}}(N,d)$, $D(N)$, and $D_{\text{a}}(d)$. For the presented experimental examples of Fig. 1 and Fig. 2, the values of the dimensions are given in Table 1. Eqs. 16 and 19 have been developed to take into account the effect of the target length N on the experimental probability distribution of synthesis products. For constant growth $D_{\text{a, measurable}}(N,d)$ is calculated by Eq. 16, whereas under non-constant multicyclic synthesis conditions we use here Eq. 19 to get exact values of $D_{\text{a, measurable}}(N,d)$. The relationship that exists between the target yield and the target length N is expressed by Eq. 20 in terms of values of $D(N)$. This is the target yield that only depends on the target length N. It can be related to the relative yield of all error sequences, that is replaceable by the target yield, in terms of $D_{\text{a, measurable}}(N,d)$ by Eq. 21. The quantities $D_{\text{a}}(d)=$

Table 1
Quantities of linear[a] multicyclic syntheses

Synthesis target length, N	d[b]	p[b]	$D_a(N,d)$[c]	$D(N)$[d]	$D_a(d)$[e]
30	$d_0 = 0.983406$	$p_0 = 0.7981075$	1.19629	0.965517	0.807090
65	$\alpha = 0.998$ $\beta = 0.00060$	$\gamma = 0.995$ $\delta = 0.00705$	1.07250	0.984375	0.917834

[a] Linear means that first of all the primary structure (sequence) has been analyzed.
[b] The parameters d_0, p_0 were automatically calculated from Eqs. 7–10 in an iterative way. The values are the means obtained from experimental data of Figs. 1A and 1B. The parameters $d_i = \alpha \exp\{-\beta(i-1)\}$ and $p_i = \gamma \exp\{-\delta(i-1)\}$ were calculated from Eqs. 13 and 14 with experimental data of Fig. 2.
[c] For constant growth calculated from Eq. 16. For non-constant growth calculated from Eq. 19.
[d] Calculated from Eq. 20.
[e] Calculated from Eq. 21.

0.807090 for the 30mer synthesis and $D_a(d) = 0.917834$ for the 65mer synthesis cover the influence of d on the target product during multicyclic synthesis. More experimental runs of 30mer up to 120mer syntheses are underway to attribute directly disturbing factors to the variation produced by noise. However, the difference between the two $D_a(d)$ values indicates that the performance (efficiency) of this 65mer chemical oligonucleotide synthesis is higher than that of the 30mer chemical oligonucleotide synthesis presented herein.

In Fig. 3 we further consider the analysis of the data vector $f(t) = \vec{A}^{(\mathrm{expl})\,\mathrm{T}}$. The vector is the collection of $N-1$ components at sampling intervals $\Delta t = 1$. In Fig. 3A we plot $f(t)$ against the nucleotide length l for the calculated elution profile of the 65mer (dC_{65}) synthesis shown in Fig. 2, whereby l is related to t according to Eq. 22. The model $f(t)$, that is experimentally verified by HPCE as far as the $N-36$ peak, does not reveal any oscillations for the dC_{65} synthesis. The shape of the solution profile of $f(t)$ is typical for a non-chaotic process. The sample autocorrelations ϱ_{r+1} that are actually available from the existing data of the dC_{65} synthesis are shown ir Fig. 3B. The statistical dependence of one observation on another is a function only of their distance r apart. Inspection of the sample autocorrelation function of Fig. 3B indicates that for this system (dC_{65} synthesis) there exists a positive autocorrelation ϱ_2 between adjacent observations, estimated by $\varrho_2 = +0.394$. Beyond this point of the serial dependence in the data the correlation pattern seems to die out. The standard error of calculation of the sample autocorrelation ϱ_{r+1} is ca. 0.125 assuming the theoretical autocorrelations are all zero. Hence, the series is said to be stationary, and in particular it has a fixed mean. The fast Fourier transformation of the dC_{65} synthesis plotted in Fig. 3C shows no characteristic frequencies ν as expected for this non-periodic process.

Next, we theoretically analyze attempted 655mer syntheses published in the literature [50]. These DNA oligonucleotide syntheses were carried out on the 1.0- or 0.2-μmol (synthesis) scale for the top and bottom strands of the synthetic *nef* gene. The 655 bp synthetic *nef* gene (EMBL accession X58780) has a 621 bp sequence which codes for a protein identical to HIV-1 (human immunodeficiency virus type 1) isolate BRU [51]. For these syntheses we suppose parameters d and p which we found for attempted 238mer syntheses [33]: $d_i = 0.940 \exp\{-0.0001(i-1)\}$, $p_i = p_0 = 0.400 = \text{const.}$ The model $f(t)$ calculated by these parameters for the 655 synthesis tends towards zero. This scenario occurs after an initial undulation. The sample autocorrelation function is nearly the same on the average at low lags r: $\varrho_2 = +0.95414$, $\varrho_3 = +0.91056$, $\varrho_4 = +0.86794$, $\varrho_5 = +0.82700$. The standard error of calculation is ca. 0.0391. The fast Fourier transformation of the 655mer synthesis indicates that a frequency at $\nu \simeq 40$ is mainly contributing to the behavior. The calculated largest Lyapunov exponents λ_1 are -0.0684600 for a two-dimensional embedding space and -0.0625392 for a three-dimensional embedding space. Because the sign of the largest Lyapunov exponents is negative, the dynamic behavior of these

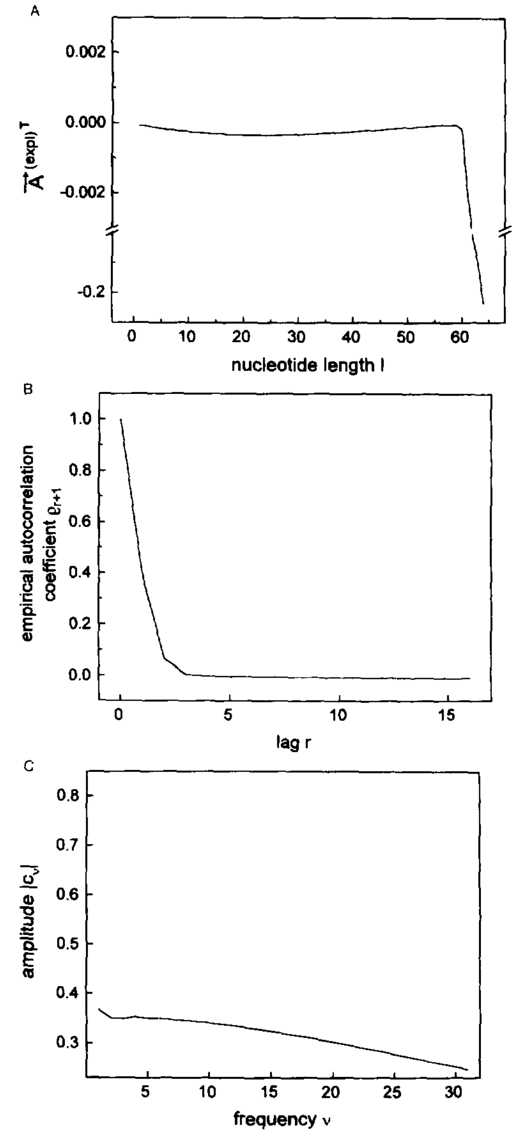

Fig. 3. Description of the measured elution profile of Fig. 2 by (A) the model function $f(t)$ of relative peak areas, (B) its autocorrelation function, and (C) its fast Fourier transformation. See text for explanation.

655mer syntheses is not chaotic. We conclude from the analysis that the solution profile $f(t)$ of these 655mer syntheses is stationary. In the particular instance of parameters d and p chosen, $f(t)$ will shrink to 0 in the limit $t \to \infty$; an elementary attractor arises.

5. Conclusions

Today the methods of polymer-supported oligonucleotide synthesis are so advanced that the product released from the solid-phase is often used without further purification. Yet, for many practical purposes, and especially in view of future applications as human therapeutics, it may be important to know, in every detail, what individual species are contained to what proportions in such a product.

The crude products of the solid-phase synthesized 30mer and 65mer oligodeoxyribonucleotides were separated by HPIEC and HPCE. In this pilot paper we have introduced mathematical methods of finding characteristic values d_0 and p_0 for constant chemical modes of growth as well as d and p for non-constant chemical modes of growth from experimentally measured elution profiles. These methods are employed by presenting the accompanying software developed. The experimental results are consistent with the theoretical models. The experimental elution profiles are used for practical reasons to evaluate the dynamics of multicylic synthesis. The main goal of this paper is to demonstrate the good correlation between our mathematical treatment and the experimental material that we have at hand. The experimental basis does not permit us to draw statistically valid conclusions as to the performance of different support systems for oligonucleotide synthesis. Further experiments are underway to predict the product patterns.

We have shown here that our theory brings a solution to a general problem of data analysis in separation techniques: We found that differences in the relative target yields (obtained from integrated absorbance of peaks) in the chromatogram/electropherogram of crude products of multicyclic synthesis account for the influence of target length N on the probability distribution. There is a real difference among relative target yields not connected with different target length N. The influence of target length N is quantified by the fractal dimension $D(N)$. From a theoretical general viewpoint, elution profiles of crude products of chemically synthesized oligoribonucleotides, oligopeptides and assembly of synthetic genes are also governed by this formalism.

Acknowledgments

We extend special thanks to Professor Dr. T.F. Nonnenmacher, Department of Mathematical Physics, University Ulm, Germany, for his continuous encouragement and for discussions. The authors thank Dr. A. Paulus, Corporate Analytical Research, Chiba, Basel, Switzerland, for providing 8% T capillaries and helpful comments. We gratefully acknowledge support for this work from Dionex Corporation, Sunnyvale, CA, USA, and Dr. D. Hauffe, Dionex GmbH, Idstein, Germany.

References

[1] L. Whitesell, D. Geselowitz, C. Chavany, B. Fahmy, S. Walbridge, J.R. Alger and L.M. Neckers, Proc. Natl. Acad. Sci. USA, 90 (1993) 4665–4669.

[2] E. Uhlmann and A. Peyman, Chem. Rev., 90 (1990) 545–584.

[3] C. Wahlestedt, E. Golanov, S. Yamamoto, F. Yee, H. Ericson, H. Yoo, C.E. Inturrisi and D.J. Reis, Nature, 363 (1993) 260–263.

[4] K.M. Takayama and M. Inouye, Biochem. Molec. Biol., 25 (1990) 155–184.

[5] G.J.M. Bruin, K.O. Börnsen, D. Hüsken, E. Grassmann, H.M. Widmer and A. Paulus, J. Chromatogr. A, 709 (1995) 181–195.

[6] W.J. Warren and G. Vella, BioTechniques, 14 (1993) 598–606.

[7] A. Chrambach, M.J. Dunn and B.J. Radola (Editors), Advances in Electrophoresis, VCH, Weinheim, 1994.

[8] C. Heller and J.L. Viovy, Appl. Theor. Electrophor., 4 (1994) 39–41.

[9] C. Heller, J. Chromatogr. A, 698 (1995) 19–31.

[10] H. Wätzig, J. Chromatogr. A, 700 (1995) 1–7.

[11] S.V. Ermakov, M.Yu. Zhukov, L. Çapelli and P.G. Righetti, J. Chromatogr. A, 699 (1995) 297–313.

[12] K. Klepárnik, M. Garner and P. Boček, J. Chromatogr. A, 698 (1995) 375–383.

[13] A. Guttman, R.J. Nelson and N. Cooke, J. Chromatogr. A, 593 (1992) 297–303.

[14] P.G. Righetti and C. Gelfi, in: P.G. Righetti (Editor), Capillary Electrophoresis in Analytical Biotechnology, CRC Press, Boca Raton, FL, 1995.

[15] A.J. Bourque and A.S. Cohen, J. Chromatogr. B, 662 (1994) 343–349.

[16] A. Belenky, D.L. Smisek and A.S. Cohen, J. Chromatogr. A, 700 (1995) 137–149.

[17] M. Vilenchik, A. Belenky and A.S. Cohen, J. Chromatogr. A, 663 (1994) 105–113.

[18] W.A. Ausserer and M.L. Biros, BioTechniques, 19 (1995) 136–139.

[19] H. Seliger, R. Bader, E. Birch-Hirschfeld, Z. Földes-Papp, K.H. Gührs, M. Hinz, R. Rösch and C. Scharpf, React. Funct. Polym., 26 (1995) 119–126.

[20] A. Usman, M. Egli and A. Rich, Nucleic Acids Res., 20 (1992) 6695–6699.

[21] A. Andrus, H. Vu, P. Ramstad and M. Pallas, Nucleic Acids Symp. Ser., 24 (1991) 41–42.

[22] J. Wyatt, M. Chastain and J.D. Puglisi, BioTechniques, 11 (1991) 764–769.

[23] B.M. Bonora, C.L. Scremin, F.P. Colonna and A. Garbesi, Nucleosides Nucleotides, 10 (1991) 269–273.

[24] S. Iwai, T. Sasaki and E. Ohtsuka, Tetrahedron, 46 (1990) 6673–6688.

[25] T. Geiser, Ann. N.Y. Acad. Sci., 616 (1990) 173–183.

[26] H. Seliger, in: S. Agrawal (Editor), Methods in Molecular Biology, Vol. 20: Protocols for Oligonucleotides and Analogs, Humana Press, Totowa, 1993, pp. 391–435.

[27] M.H. Caruthers, G. Beaton, J.V. Wu and W. Wiesler, Methods Enzymol., 211 (1992) 3–20.

[28] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York, 1983.

[29] Z. Földes-Papp, E. Birch-Hirschfeld, S. Conrad, B. Möpps, H. Seliger and A.K. Kleinschmidt, (1996) in preparation.

[30] Z. Földes-Papp, A. Herold, H. Seliger and A.K. Kleinschmidt, in: T.F. Nonnenmacher, G.A. Losa and E.R. Weibel (Editors), Fractals in Biology and Medicine, Birkhäuser, Basel, 1994, pp. 165–173.

[31] R.L. Devaney, An Introduction to Chaotic Systems, Benjamin/Cummings, Menlo Park, CA, 1986, pp. 39–43.

[32] Z. Földes-Papp, E. Birch-Hirschfeld, R. Rösch, M. Hartmann, A.K. Kleinschmidt and H. Seliger, J. Chromatogr. A, 706 (1995) 405–419.

[33] Z. Földes-Papp, W.-G. Peng, H. Seliger and A.K. Kleinschmidt, J. Theor. Biol., 174 (1995) 391–408.

[34] R.P. Iyer, D. Yu, Z. Jiang and S. Agrawal, Nucleosides Nucleotides, 14 (1995) 1349–1357.

[35] B.B. Mandelbrot, in: T.F. Nonnenmacher, G.A. Losa and E.R. Weibel (Editors), Fractals in Biology and Medicine, Birkhäuser, Basel, 1994, pp. 8–21.

[36] B.J. West and W. Deering, Phys. Rep., 246 (1994) 1–100.

[37] B. Abraham, and J. Ledolter, Statistical Methods for Forecasting, Wiley, New York, 1983, pp. 60–74.

[38] R. Seydel, Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos, Interdisciplinary Applied Mathematics, Vol. 5, Springer, New York, 1994, 2nd ed., pp. 340–357.

[39] N.H. Packard, J.P. Crutchfield, J.D. Farmer and R.S. Shaw, Phys. Rev. Lett., 45 (1980) 712–715.

[40] F. Takens, in: D.A. Rand and L.S. Young (Editors), Dynamical Systems and Turbulence, Lecture Notes in Mathematics, Vol. 898, Springer, Berlin, 1981.

[41] E. Birch-Hirschfeld, Z. Földes-Papp, K.-H. Gührs and H. Seliger, Nucleic Acids Res., 22 (1994) 1760–1761.

[42] E. Birch-Hirschfeld, Z. Földes-Papp, K.-H. Gührs and H. Seliger, Helv. Chim. Acta, 79 (1996) 137–150.

[43] T. Manabe, N. Chen, S. Terabe, M. Yohda and I. Endo, Anal. Chem., 66 (1994) 4243–4252.

[44] G. Baumann and Z. Földes-Papp, Wolfram Research, IL, USA, Mathsource, 1996.

[45] A. Wolf, J.B. Swift, H.L. Swinney and J.A. Vastano, Physica, 16D (1985) 285–317.

[46] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, Numerical Recipes in Fortran: The Art of Sientific Computing, Cambridge University Press, New York, 2nd ed., 1992.

[47] K.L. Fearon, J.T. Stults, B.J. Bergot, L.M. Christensen and A.M. Raible, Nucleic Acids Res., 23 (1995) 2754–2761.

[48] Y.-Z. Xu and P.F. Swann, Anal. Biochem., 204 (1992) 185–189.

[49] J. Temsamani, M. Kubert and S. Agrawal, Nucleic Acids Res., 23 (1995) 1841–1844.

[50] R.B. Cicarelli, P. Gunyuzlu, J. Huang, C. Scott and F.T. Oakes, Nucleic Acids Res., 19 (1991) 6007–6013.

[51] R. Weiss, N. Teich, H. Varmus and J. Coffin, RNA Tumor Viruses, Vol. 2, Cold Spring Harbor Laboratory Press, New York, 2nd ed., 1985, Appendix 3, pp. 1102–1123.