Applying signal theory to the analysis of biomolecules

Gerhard Kauer* and Helmut Blöcker

Department of Genome Analysis, GBF—German Research Centre for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Received on November 21, 2002; revised on March 27, 2003; accepted on May 3, 2003

ABSTRACT

Motivation: The accumulation of sequence-related and other biological data for basic research and application purposes invites disaster. It appears very likely that neither traditional thinking nor current technologies (including their foreseeable evolutionary developments) will be able to cope with this ever intensifying situation.

Results: We present the detailed theoretical background for applying signal theory, as known from speech recognition and image analysis, to the analysis of biomolecules. The general scheme is as follows: biochemical and biophysical properties of biomolecules are used to model an \( n \)-dimensional signal which represents the entire information-bearing biomolecule. Such signals are used to search for biological principles, analogies or similarities between biomolecules. In a series of simple experiments (bacterial DNA, generation of real signals using melting enthalpies, detection filtering by convolution of signals) we have shown that the novel system for comparative analysis of the properties of information-bearing biomolecules works as in theory.

Contact: GerhardKauer@t-online.de

Supplementary information: http://genome.gbf.de/wavepaper

INTRODUCTION

Gene-based research and development have opened a new era in biology, biomedicine and biotechnology. This has led to entirely new challenges for the IT sector. The storage capacity for a single human genome alone is in the order of dozens of gigabytes. In the near future we will see more data about the dynamics of living systems and their molecules as opposed to the current data which vastly describe the statics of the biological systems. On the other hand, the information behind static data like, for example, the primary structure of DNA represents another mining treasure which will contribute to the foreseeable data disaster in gene research. It is commonly believed that the amount of data in gene research doubles every six months.

This is much faster growth than, for example, the number of circuits on a microchip which, according to Moore’s law, doubles only every 18 months. Speeding up the clock rate of computer processors and assembling masses of traditional hardware alone does not solve the problem. As a consequence, major IT players have announced specific design and engineering of next generation hardware to have a chance to be able to meet the specific demands of gene research. Even if the optimized hardware was at its best, the software currently employed to the analysis of information carrying biomolecules is not. State-of-the-art here is character-based search and pattern analysis. The information behind the primary structure (regarding three-dimensional information like loop building and its influence on transcription, enhancer/silencer effects) of DNA for example is too complex for these relatively inflexible algorithms. Another important challenge is the fragment assembly of sequenced DNA as in whole genome shotgun sequencing strategies. Current methods take rather long cpu times and often result in suboptimal assemblies.

SYSTEMS AND METHODS

To develop entirely new software approaches for the analysis of information carrying biomolecules like DNA or proteins, we decided to pursue a different path. We connect the proven methodology of image analysis and speech recognition to one of the most challenging problems in modern biosciences. The first two deal with complex and varying data sets and hidden information on a rather sophisticated level. A wide field of application with a big arsenal of methods is thus available for analysis of information carrying biomolecules.

In principle, it can be stated that biological macromolecules such as DNA and proteins are not present in the cell as sequences of mnemonic letters. They rather are chemical substances connected to their immediate environment through a multitude of chemical and physical interactions down to the atomic level. These effects are not only to be found on the primary structure level, but go far beyond this level. Spatial structures, charge and enthalpy data also play a role when these information carriers of life fulfil their tasks. It appears
desirable to also include these data as interpretable signals in mathematical models in addition to pure sequence comparisons in order to analyze the ‘sequence function’ of these macromolecules. Large-scale genomic comparisons could also benefit from a mathematical modeling of the problem. Thus, for example, more extensive stochastic analyses on this level might reveal interesting insights into evolution, taxonomy and functionality of DNA.

In view of these prospects, methods of signal theory seem to have been made for the analysis of biomolecules. The critical point would only be to find a way of transforming the traditional biological codes (for example letter code) into a signal, which is an indispensable prerequisite for this type of investigation.

Finding relevant data sets

It would be completely wrong to transform e.g. the nucleotides of DNA into a series of arbitrarily chosen data (e.g. \( A = 1, G = 2, T = 3, C = 4 \)). Such an assignment would not provide a real ‘signal’ (Karu, 1995). No ‘similarity’ as we understand it in biological research will thus be obtained, because there is no valid physical or chemical relation behind it. Initial transformation experiments were therefore based on DNA enthalpy data describing the melting of DNA double strands. It should be noted here, that choosing enthalpy data is only one out of many possibilities of achieving a signal from a DNA primary structure. We have chosen this example, because of its clearness.

The enthalpy data are dependent on the respective neighboring nucleotide and can be taken from the literature (Breslauer et al., 1986). In some cases, however, these experimental enthalpy data are identical so that their unequivocal assignment is not possible. To overcome this obstacle, the accuracy was extended by the rounding error. For example, the enthalpy data provided from the literature for the nucleotide pairs CA and TG is 5.8 and was extended to CA = 5.75 and TG = 5.84. This extension (‘reverse rounding’) simply serves as mathematical differentiation, maintaining the physical values as in the literature. The physical signal characteristics are not changed by this procedure (Sachs, 1982).

The sequence data are transferred into numerical data using Table 1. In doing so, it must be ensured that the primary sequence is only transformed in the 5’ \( \rightarrow \) 3’ direction (leading strand). The reason is that the enthalpy data are not only attributable to the direction-invariant hydrogen bonds between the single strands but also to the interactions between the electrons of the respective neighbouring base pairs (hydrophobic base stacking). Moreover, van der Waals forces contribute not insignificantly to the interactions between the immediate base neighbors. This information is not reversible for the strand direction and must therefore be taken into account in the enthalpy-based conversion of the primary structure into a signal (Chalikian et al., 1999). See simple example (Fig. 1).

### Table 1. Values used to transform a DNA primary structure into a signal

<table>
<thead>
<tr>
<th>Nucleotide pair</th>
<th>( \Delta H_a^b ) (kcal/mol)</th>
<th>( \Delta H_b^b ) (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’ ( \rightarrow ) 3’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>9.1</td>
<td>9.05</td>
</tr>
<tr>
<td>AT</td>
<td>8.6</td>
<td>8.60</td>
</tr>
<tr>
<td>TA</td>
<td>6.0</td>
<td>6.00</td>
</tr>
<tr>
<td>CA</td>
<td>5.8</td>
<td>5.75</td>
</tr>
<tr>
<td>GT</td>
<td>6.5</td>
<td>6.45</td>
</tr>
<tr>
<td>CT</td>
<td>7.8</td>
<td>7.75</td>
</tr>
<tr>
<td>GA</td>
<td>5.6</td>
<td>5.55</td>
</tr>
<tr>
<td>CG</td>
<td>11.9</td>
<td>11.90</td>
</tr>
<tr>
<td>GC</td>
<td>11.1</td>
<td>11.10</td>
</tr>
<tr>
<td>GG</td>
<td>11.0</td>
<td>10.95</td>
</tr>
<tr>
<td>TG</td>
<td>5.8</td>
<td>5.84</td>
</tr>
<tr>
<td>AC</td>
<td>6.5</td>
<td>6.54</td>
</tr>
<tr>
<td>AG</td>
<td>7.8</td>
<td>7.84</td>
</tr>
<tr>
<td>AG</td>
<td>7.8</td>
<td>7.84</td>
</tr>
<tr>
<td>TC</td>
<td>5.6</td>
<td>5.64</td>
</tr>
<tr>
<td>CC</td>
<td>11.0</td>
<td>11.04</td>
</tr>
<tr>
<td>TT</td>
<td>9.1</td>
<td>9.14</td>
</tr>
</tbody>
</table>

\( ^{a} \)As taken from (Breslauer et al., 1986).

\( ^{b} \)For detection filtering the values for NN, NA, NG, NT, NC, AN, GN, TN, CN have to be set to zero, for other degenerate codes we propose to use mean values.

### Fourier transformation

One attribute of information-carrying biomolecules like DNA or protein is, that data sets associated to them are really large-scale. They may easily reach billions of single data points. Hence, the signal derived from them is often very complex. Since ‘speed matters’, this fact must be considered for setting up a competitive technology.

Complex measurement data can be transferred into a mathematically much more easy-to-handle sum of sine and cosine functions via Fourier (Brigham, 1974, www.catalina-research.com) transformation (polynomes, Fig. 1 at http://genome.gbf.de/wavepaper). It should be noted here that other methods to process signals including wavelet analysis, fuzzy logic and transformation analyses after Hartley and Hough may also be applied. The Fourier transform is a function which can indicate the amplitude and phase of the sine function associated with each frequency. The amplitude describes the deflection of the sine function, whereas the phase defines the function’s passage through the ordinate origin (for a detailed description of the mathematical background see http://genome.gbf.de/wavepaper). Fourier transformation of complex data sets, accelerated by special DFFT signal processors, represents an efficient analysis platform. Furthermore, since we deal with real numbers, combined Fourier transformations (DFFT/DIFFT) are possible. After their transformation many mathematical operations (e.g. in digital image analysis and speech recognition) can be applied much more easily and rapidly. The methods applied to speech recognition, for example, involve algorithms to filter energy contents,
patterns or conspicuous structures out of signals and hence they become identifiable.

DNA and protein analyses may be regarded as a problem very similar to digital image analysis and speech recognition. After transforming strings of biological code into signal energy, the aim is to find, on the signal energy level, relevant motifs on DNA and protein macromolecules, to compare them with each other and/or analyze them with respect to their biological significance.

Filter methods

The general steps in this methodology are (i) finding a relevant data set for generating signals, (ii) producing the signals, (iii) signal analysis (in the frequency domain, in the spatial domain, by wavelet analysis). Any filter has to be designed depending on the logic of steps (i)–(iii). For the comparative analysis of nucleic acid sequences we have chosen to use nearest neighbor analysis-based melting enthalpy data to enable single nucleotide resolution [steps (i) and (ii)]. A rather simple mathematical filter method to compare such signals is convolution and the sibling method of correlation (Press et al., 1988). It is used as ‘detection filter’ in image analysis (Wahl, 1989), where it is a well-known method for structure recognition. The detection filter is based on the assumption that the signal energy of a structure searched for appears again in the target structure. In the spatial domain this discrete correlation must be calculated by

$$C(u, v)_i = \sum_{j=0}^{N-1} u_{i+j} h_j$$

whereas in the frequency domain it is merely a multiplication

$$C(u, v)_i = U^*_j V_j$$

where $U, V$ are the Fourier transforms and $^*V$ denotes the conjugate complexes of $V$.

This massive simplification of the correlation filter is why, for the comparison of DNA sequences, we decided to move into the frequency domain using the discrete Fourier transformation. Detection filtering in the present case is done by multiplication of the conjugate complex Fourier transform of the search signal by the Fourier transform of the target signal.

Acceleration

If it is intended to convolute large data volumes in this way via Fourier transformation, this may cause severe computer performance problems. Methods of Fourier transformation, such as discrete fast Fourier transformation (DFFT), are based on sorting algorithms which depend on the simultaneous presence of all data of a given data set. Analyses of the human genome are in the range of $n$-fold $10^9$ data points, considering the primary sequence alone. It is therefore desirable to segment the extensive data set so that, on the one hand, an application of DFFT as efficient as possible is ensured and, on the other hand, the enormous data volumes can be analyzed via these segments without electronic storage limitation. To exclude any disturbance of the signal by the segmentation procedure, the well established method of overlap add segmentation is applied (Brigham, 1974).

In a pure software implementation of the method as detailed here, Fourier transformation is the rate-determining step of the analyses, in spite of any software acceleration by the application of methods like overlap add segmentation. However, acceleration of Fourier transformation by
hardware support is also possible. Due to its general technical significance, special digital processors have already been developed. These accelerators are now available as commercial solutions. With such an equipment, approximately 4.7 GFLOP operations (5,242,880 operations / 1098.25 μs) are possible. The commercially available ‘Cheetah’ PCI card enables 4.774 GFLOPS doing a 64K complex FFT (www.catalina-research.com), so that corresponding data volumes can be processed using, again, the overlap add segmentation method. For the human genome we would like to present the following scenario: transformation method: Fourier (please note: alternatives are at hand!), cpu: Pentium 4 or similar, hardware acceleration (DSP as described above).

Following this scenario and taking into account about 3000 overlap add segmentations, all transformations (DFFT and DIFFT) would take about 5.3 s. This represents the only ‘bottleneck’ of the technology as described in this manuscript. We would like to stress here again that the speed is widely independent of the number and complexity of queries.

For the future application of our general approach to complex tasks like modeling of cellular infection processes, the acceleration perspectives as described above may not be sufficient. To reduce the speed limitations by the transfer rates from hard disks through the PCI bus, for example, bit-coded tables may be used. Optionally, these tables may be compressed.

Furthermore, another possibility to accelerate the method would be to rely on a purely electronic solution rather than hard disks as the storage medium for time-critical routines. In view of the data volumes to be analyzed and decreasing memory prices, this may represent a realistic option for the future.

Initial experiments

Motif-detection is a main stream problem in hundreds of papers [see e.g. reviews by Stormo (2000) or Jiang et al. (2002)]. Simple sequence comparisons with our technology are already possible by just transforming sequence data into a signal. As a first test case we decided to search for fragments of the form CNNGAANNTCNNG (heat shock sequence motif) in a number of bacterial genomes. The data obtained from the signal generation of search and target sequences were Fourier-transformed and then the conjugate complex of the search signal was formed. In a further step, the signal energy of the search structure was determined by multiplying the conjugate complex Fourier transform of the search signal by the Fourier transform of the search signal (autocorrelation). The retransformation of this result provided the significant signal energy of the search structure as the maximum value. Normalization could be effected through the variances (deviation squares of the amplitudes, summed averages). In a further step, the conjugate complex Fourier transform of the search structure was multiplied by the Fourier transform of the target structure and the result was subjected to discrete inverse Fourier transformation. Finally, the absolute position of the sequence searched for is found by localizing the above-calculated signal energy of the search structure in the result (Fig. 1).

All motifs (Ron et al., 1999) of the target sequences could be found. This shows that the general idea of transforming primary structures of DNA into real signals and applying the mathematics of signal theory to them leads to successful search results. For a 5 Mb bacterial genome the entire process took less than 3 min on a Sun Blade 1000 not involving any hardware acceleration as described above.

It is noteworthy that the simple test case (correlation filter) already shows an additional potential of the technology. On top of spotting sequences of the general structure of the search sequence, additional information about related sequences is provided (Figs 2 and 3) by setting an appropriate window of the signal energy.
Correlation (%): 100.0. (A) C TAGC ACT C N N N N N N NG A G T G C T A G
(B) C GAGC ACT C N N N N N N N A G G C C C A G
(C) C N AGC ACT C N N N N N N N A G N G C N A G

Fig. 3. Influence of artificial mutations in the outer cassettes of the heat shock motif (GroEL2) on search results in M. tuberculosis (Ron et al., 1999). The unmodified motif is easily detected starting at position 528,606 of the genome. To demonstrate the possibilities of selectivity and sensitivity of our method, the target sequence is manipulated on some nucleotide positions (in bold). The sequence is therefore different from the motif to be found and the relative correlation declines to 83%. Using Ns on the positions of mutation, the motif can again be found with 100% correlation. Heat shock motif: (A), Mutation: (B), Masked query: (C), Energy (A, B): 1016.0, Autocorrelation (B, B): 1221.5, Relative Correlation (%) = 83.2%, Energy (A, C): 629.3, Autocorrelation (C, C): 629.3, Relative Correlation (%) = 100.0%, Energy (A, C): 629.3, Relative Correlation (%): 100.0.

DISCUSSION AND CONCLUSION

In a series of simple experiments (bacterial DNA, generation of real signals using melting enthalpies, detection filtering by convolution of signals) we have shown that the novel system for comparative analysis of the properties of information-bearing biomolecules works as in theory. There is basically no length limitation, neither for the target nor for the search sequence. Sensitivity and selectivity can be easily adjusted at will. The system as it stands is ready for use with the longest eukaryotic genomes. The generation of the signals is of crucial importance. The choice of the physico-chemical data sets, which are used to generate signals, determine the field of application of the methodology. For example, using melting enthalpies enables one to search for virtually any pattern in nucleic acids down to the nucleotide resolution. It also appears interesting to investigate long range regularities in genomes by applying suitable (for example low-pass) filters. An intriguing example may be the search for functional gene domains in eukaryotic genomes through the analysis of stress induced duplex destabilization (SIDD) (Benham et al., 1977). An appealing future application of high resolution pattern detection in DNA sequences may lie in high speed and high quality whole genome shotgun assembly.

To this end, we have successfully carried out numerous experiments (data not shown) similar to those in the previous section (further consensus sequences of DNA-binding proteins and restriction sites etc.). Given an appropriate data set to generate signals from polypeptide chains (based on the physicochemical properties of the amino acid side chain, codon usage, wobble base analysis, secondary structure motifs etc.), proteins can be investigated following the same scheme as for nucleic acids and similarly any other information-carrying macromolecule. It should be noted that signal analysis is per se n-dimensional. In our current investigations we have exclusively performed one-dimensional analyses (transformed primary structure). But as it is already standard in other fields of applied signal theory, one can imagine analyses of multi-dimensional signals in the field of biomolecules, too. However, all components of the signal vector must be strictly correlated by their chemical or physical properties. Otherwise, by its definition, a signal would not exist. Such more-dimensional analyses could be used to find analogies rather than mere homologies. Based on the technology as presented here one may consider, for example, revisiting the analysis of promoters and splice sites. Our approach can detect similarities which do not become apparent by looking at letter codes. One major point of our paper is the fact that we base any analysis on physicochemical properties. Therefore, nucleotide sequences which do not even show remote homology by their letter codes may nevertheless show similar signal characteristics of potential biological significance. And we feel this could be of some benefit in the long run for revisiting the understanding of promoters or other regulatory elements. In fact, this is part of our current research. Since such analyses would be based on an entirely new approach, unlike weight matrices or bare statistics, one may expect new clues on the functional impact of these elements.

The general strategy as outlined here allows for a number of alternative technical approaches. While small data sets (up to about 100 kb of transformed primary structure) may be best analyzed in the spatial domain, larger data sets will preferably be processed in the frequency domain by applying, for example, Fourier transformation. Recently, we have started to work on wavelet analysis as a further alternative.

In summary, we have only begun to explore the potential of the application of signal theory to the analysis of biomolecules. To this end, we have shown in simple experiments that much faster and cheaper procedures can be developed than following traditional routes and also that new questions can be tackled. We would appreciate joining forces with others to increase the power of the methodology and to conquer new fields of application like modeling of infection processes and systems biology.
ACKNOWLEDGEMENTS

G.K. was supported by a grant from Deutsche Forschungsgemeinschaft (KA 1655/1-1) and a grant from German Ministry for Education and Research (BMBF) through PII (O31U210A).

REFERENCES


