FeatureScan: revealing property-dependent similarity of nucleotide sequences

Igor V. Deyneko^{1,2,*}, Björn Bredohl^{1,4}, Daniel Wesely^{1,4}, Yulia M. Kalybaeva¹, Alexander E. Kel³, Helmut Blöcker^{1,*} and Gerhard Kauer^{1,4}

¹Department of Genome Analysis, GBF (German Research Centre for Biotechnology), D-38124 Braunschweig, Germany, ²Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, ³BIOBASE GmbH, D-38300 Wolfenbüttel, Germany and ⁴University of Applied Sciences, D-26723 Emden, Germany

Received February 14, 2006; Revised March 1, 2006; Accepted April 17, 2006

ABSTRACT

FeatureScan is a software package aiming to reveal novel types of DNA sequence similarity by comparing physico-chemical properties. Thirtyeight different parameters of DNA double strands such as charge, melting enthalpy, conformational parameters and the like are provided. As input FeatureScan requires two sequences, a pattern sequence and a target sequence, search conditions are set by selecting a specific DNA parameter and a threshold value. Search results are displayed in FASTA format and directly linked to external genome databases/browsers (ENSEMBL, NCBI, UCSC). An Internet version of FeatureScan is accessible at http://genome.gbf.de/featurescan/. As part of the HOBIT initiative (http://hobit.sourceforge.net/) FeatureScan is also accessible as a web service at its above home page. Currently, several preloaded genomes are provided at this Internet website (Homo sapiens, Mus musculus, Rattus norvegicus and four strains of Escherichia coli) as target sequences. Standalone executables of FeatureScan are available on request.

INTRODUCTION

The principle of similarity measurement lies behind any recognition method or sequence analysis. Motifs, weight matrices, Markov models significantly differ in mathematical background (1–4), but they all rely on more or less sophisticated statistics of four mnemonic letters (A, T, G, C). Even the most complex of such methods are still not able to specifically and accurately solve the problem of detecting regulatory DNA sequences.

FeatureScan is based *ab initio* on a different principle. It works with numerical sequences which describe specific properties of DNA and utilizes methods from signal theory (5) to compare them. One convincing argument of many in favour of such an approach is that it has been observed that the binding sites for the HFN1 transcription factor require its own specific melting characteristic of the surrounding region (6). For an appropriate modelling of this fact, it appears adequate to consider melting enthalpy of DNA rather than the bare alphabetic sequence.

METHODS

Algorithmically, FeatureScan originates from proven methodologies in image analysis (7) and speech recognition (8). The current implementation is based on a convolution method (9) and can be described briefly in three main steps (for the detailed theoretical background see our earlier publication (10)).

First is a transformation of nucleotide sequences (pattern and target sequence) into numerical form, which we refer to as signals. At this step users have to decide which property may play an important role in their specific cases [see (11) and below].

Second is a computation of the correlation integral (1) of two signals f and g, which can be rewritten using Fourier transformants F and G yielding (3). Assuming having direct and inverse Fourier transformations implemented (in our case it is optionally hardware implemented), this step is reduced to just a multiplication. The final step is looking for shift values y which will define possible matches of the sequences. The difference between correlation (3) and

*To whom correspondence should be addressed. Tel: +49 531 6181 224; Fax: +49 531 6181 292; Email: ide@gbf.de *Correspondence may also be addressed to Helmut Blöcker. Tel: +49 531 6181 220; Fax: +49 531 6181 292; Email: bloecker@gbf.de

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

autocorrelation (2) integrals must be less than the predefined threshold.

$$\operatorname{Corr}(y) = \int f(x) \cdot g(x - y) \, \mathrm{d}x, \qquad 1$$

AutoCorr =
$$\int g(x) \cdot g(x) dx$$
, 2

 $Corr(y) = InverseFourierT\{F(y) \cdot \overline{G(y)}\}.$ 3

PROGRAMME/WEBSITE DESCRIPTION

FeatureScan is available as a local standalone application and via Internet, both having equivalent core functionality. The local, command-line version lends itself to extensive batch processing. The web service (following HOBIT standards, http://hobit.sourceforge.net/) offers an additional advantage: using the optional hardware acceleration (fast Fourier transformation PowerFFT card from Eonic, www.eonic.com) boosts the entire procedure up to 470 times compared with a 1.7 GHz AMD processor PC alone. This allows scanning, for example, through the entire human genome in ~ 3 min. The core programme is written in compliance with ANSI C standards and can be compiled on many platforms [tested under Windows, Linux (RedHat, SuSe), FreeBSD]. Currently, there are different implementation-dependent limitations with respect to the maximum length of the sequences (see FeatuteScan help page).

To run FeatureScan via Internet, users have to provide a pattern sequence (field 1 in Figure 1a) and a target sequence (field 2 in Figure 1a). All sequences must be in one of the following formats: EMBL, FASTA or plain text. A target sequence can either be uploaded or selected from the preloaded human, mouse, rat or Escherichia coli genomes. A pattern sequence is assumed not to be longer than a target sequence. In field 3 (Figure 1a) it is specified which of the 38 DNA parameters will be used in the analysis. A threshold value needs to be entered in field 4 (Figure 1a) which defines the stringency of the analysis run. It must be noted here that due to the entirely different theoretical background the similarity values are not always suited for a 1:1 comparison to BLAST-type similarity values [for further details see (11)]. For a first brief inspection of the programme the user may wish to choose our example.

An important step of the procedure is selecting a suitable property. As mentioned before, 38 parameters of DNA double strand, comprising physico-chemical (melting temperature, entropy and others) and conformational (roll, tilt, slide, twist and others) DNA characteristics, have been collected in a public database (6). A brief overview of a few of the parameters is given in Table 1 and of all of them on the FeatureScan help page. It should be pointed out that the datasets cannot be ordered by their recognition power (by selectivity, false positive rate and others) to be able to recommend one universally accepted best parameter. In numerous examples it has been demonstrated that various functions of different DNA *loci* are implicated by various DNA physical properties (6). Thus, we do not recommend

any 'best-recognizing' parameter to the user. Instead, we suggest the user to speculate about the possible background of his specific task and then choose the most relevant parameter.

As a last rescue, we would recommend to start with *'MeltEnthalpyBreslauer'*, because melting often plays a role in DNA processing.

Results of the search are either displayed in a table or in FASTA format, which makes further analysis more comfortable (Figure 1b). The name line (following '>') consists of consecutive number, similarity value and starting position of a matched subsequence in the submitted target sequence. For pre-loaded genome sequences, links to the corresponding sequences in the external databases are also added to the name.

FeatureScan WEB SERVICE

To take full advantage of the accelerating power of the FFT card and to avoid the tedious routine of typing sequences and parameters on an Internet page, the FeatureScan web service was developed. Such Internet services allow user programmes to directly call FeatureScan on a remote machine as if it was a local function. The key point of such web services is an interface (between user and server programmes), the standardized description of the number and format of parameters which are passed on from the user programme to FeatureScan and a returned result. Our implementation of the interface complies with the rules of the HOBIT initiative (http://hobit.sourceforge.net/).

The FeatureScan web service is implemented in a Client– Server architecture. Any user client programme set up according to the rules can send a request to the database, and the job will be tagged with a unique identifier. The server programme connects to the database, picks up a task from the queue and returns a result. By querying the database, the user programme will track the status of the task and may pick up results. A client application, written in C++ (Windows), can be downloaded from the FeatureScan web page (multiplatform Java client to follow soon).

The advantages of such an architecture lies in its flexibility and scalability. Both, user client programme and server programme may be developed independently to meet specific demands, advanced functionality or improved performance. To add on to the overall existing performance one has only to run another copy of the server application. Currently, two server programmes are running in Braunschweig and in Emden (Germany).

VALIDATION

To demonstrate aspects of the nature of this novel similarity measure we carried out a series of experiments with artificial and genomic sequences as published in Refs (11,13). Tests on randomly generated data showed that the method performs as in theory. It is robust to interspersed 'noise' nucleotides, able to detect complex multi-sequence elements, has single nucleotide resolution and is easily applicable to genome-wide analysis. Thirty-eight different DNA parameters show a wide range of sensitivities *versus* letter mismatches, obviously

C		iarks Loois Window Help	abf de/cai.bio/Easter-C-	an/EestureScan ol	
0		http://gnaweb		an/Featurescan.pl	Eachbochschule BIOD
G	BF	KUT	Fea	Ture & Sean	Users of Azeled Scenes Orthograde Volternaturen
			searchi by comparir	ng for sequence similarity ng physico-chemical properties of DNA	
					(1
1. Er	nter the pattern s	equence you want to	o search for by ty	ping or by copy&paste (<u>plain</u> ,	
>CSD	HR51 standar	d; DNA;	TTGAGGAGTCGTCCTC	TACCTTCTCTCCTCCCTCCCTCCCCTCCCCCCCC	
GCGG	TGGATCTCAGGCTTC	CGGAAGACTGGAAGAAC	GGCTCAGAACCGCTT	GTCTCCGCGGGGCTTGGGCGGCGGAAGA	
2. Er copy	nter the target sea &paste (plain, F.	quence in which you ASTA or EMBL for	1 want to search f rmat, check lengtl	or the pattern by typing or by h limitations): O	or choose a preloaded genom
					C Escherichia coli K12
					3 chr homo sapiens
					C 1 chr mus musculus
					C 1 chr rattus norvegicus
					CHECK genomes versions
3. Se 4. Sp	lect DNA param ecify threshold v	eter: 0 MeltEnthalpyBri rahue ($0 \le \text{Thr} \le 1$)	eslauer 0.05 (normalized	€ (Brief parameters d l difference, threshold=0 corresponds to	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specif Or che (read	lect DNA param ecify threshold v ÿ output format: Fas eck out our example our "how to start" tip	eter: $\begin{bmatrix} 0 \text{ MeltEnthalpyBridge} \\ \text{ralue } (0 \le \text{Thr} < 1) \end{bmatrix}$ sta (default, recommend :: pre-selected pattern, pos).	eslauer 0.05 (normalized ed) C or Table-viet parameters and <i>E. co</i>	☑ (Brief parameters d I difference, threshold=0 corresponds to w [©] th K12 genome as target sequence: [©]	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specif Or che (read	lect DNA param ecify threshold v ý <u>output format</u> : Fas eck out our example our "how to start" tip	eter: $\begin{bmatrix} 0 \text{ MeltEnthalpyBr}\\ \text{ralue } (0 \le \text{Thr} < 1) \end{bmatrix}$ sta (default, recommend :: pre-selected pattern, p os).	eslauer 0.05 (normalized ed) C or Table-vie parameters and <i>E. co</i>	(Brief parameters d difference, threshold=0 corresponds to w © di K12 genome as target sequence: C Submit Ouery Reset	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specif Or che (read	lect DNA param ecify threshold v ŷ <u>output format</u> : Far eck out our example our "how to start" tip	eter: $\begin{bmatrix} 0 \text{ MeltEnthalpyBridge} \\ \text{ralue } (0 \le \text{Thr} < 1) \end{bmatrix}$ sta (default, recommend :: pre-selected pattern, pos).	eslauer 0.05 (normalized C or Table-vier parameters and <i>E. co</i>	(Brief parameters d I difference, threshold=0 corresponds to w © <i>th K12</i> genome as target sequence: © Submit Query Reset Results page	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specifi Or che (read -	lect DNA param ecify threshold v ŷ output format: Fas eck out our example our "how to start" tip	eter: 0 MeltEnthalpyEntrahue ($0 \le Thr < 1$) sta (default, recommend : pre-selected pattern, press).	eslauer [0.05] (normalized ed) C or Table-vie parameters and <i>E. co</i>	(Brief parameters d difference, threshold=0 corresponds to w © di K12 genome as target sequence: C Submit Query Reset Results page	escription can be found on the <u>help page</u>) perfect matches)
3. Sei. 4. Sp Specif Or ché (read	lect DNA param ecify threshold v fy <u>output format</u> : Far eck out our example our "how to start" tip progress bar	eter: 0 MeltEnthalpyBr ralue ($0 \le Thr < 1$) sta (default, recommend :: pre-selected pattern, p os).	eslauer 0.05 (normalized C or Table-vie parameters and <i>E. co</i>	☑ (Brief parameters d I difference, threshold=0 corresponds to w ☑ th K12 genome as target sequence: ☑ Submit Query I Reset Results page	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specif Or che (read : Your 	lect DNA param ecify threshold v fy output format: Fas eck out our example our "how to start" tip progress bar	eter: 0 MeltEnthalpyBri rahue ($0 \le Thr < 1$) sta (default, recommend :: pre-selected pattern, p $\frac{250}{2}$).	eslauer [0.05] (normalized ed) C or Table-vie parameters and <i>E. co</i>	(Brief parameters d I difference, threshold=0 corresponds to w © <i>li K12</i> genome as target sequence: Submit Ouery Reset Results page	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specif Or chư (read - Your 	lect DNA param ecify threshold v ÿ output format: Fas eck out our example our "how to start" tip progress bar Ready shold=0.050000 coding scheme u	eter: $\begin{bmatrix} 0 \text{ MeltEnthalpyBrit}^{\text{Brit}} \\ \text{ralue } (0 \leq \text{Thr} < 1) \\ \text{sta (default, recommend} \\ \text{:: pre-selected pattern, p} \\ \text{sec} \\ \text{sec} \\ \text{sec} \\ \text{MeltEnthalpy} \\ \text{sec} \\ \text{MeltEnthalpy} \\ \text{sec} \\ \text{MeltEnthalpy} \\ \text{meltEnthalpy} \\ \text{sec} \\ \text{MeltEnthalpy} \\ m$	eslauer 0.05 (normalized C or Table-vier parameters and <i>E. co</i>	(Brief parameters d I difference, threshold=0 corresponds to w w (hi K12 genome as target sequence: Submit Query Reset Results page	escription can be found on the <u>help page</u>) perfect matches)
Your Your Three DNA c Your Specif	lect DNA param ecify threshold v fy output format: Far eck out our example our "how to start" by progress bar Ready shold=0.050000 roding scheme u pattern: DRRSA stand GGGATCTCAGGCTTC	eter: 0 MeltEnthalpyBr rahue (0 ≤ Thr < 1) sta (default, recommend : pre-selected pattern, p 25). sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGGGGCCT CGGAAGACTGGAAGAAC	eslauer 0.05 (normalized C or Table-vie parameters and <i>E. co</i> PBreslauer TGAGAAGTOGTCCTC GGCTCAGAACGGTT	(Brief parameters of I difference, threshold=0 corresponds to w © th K12 genome as target sequence: © Submit Ouery Reset Results page TACCTTCTCTGCTGGCTGGGGGGGAC GTTTCCCCGGGGCTTGGCTGGGGGGGAC	escription can be found on the <u>help page</u>) perfect matches)
Your Your Threes DNA c Your CCGCJ GCGGT	lect DNA param ecify threshold v ÿ output format: Fas eck out our example our "how to start" tip progress bar Ready shold=0.050000 soding scheme u pattern: 3DRSA stand AGCCAAGGCCGACGG rGGATCTCAGGCTTC	eter: OMeltEnthalpyEn ralue (0 ≤ Thr < 1) sta (default, recommend : pre-selected pattern, p se). sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGAGAGACC CGGAAGACTGGAAGAACC chr3, length=19950	eslauer [0.05] (normalized ed) C or Table-vie parameters and <i>E. co</i> /Breslauer restaceAcroencerce coccectAgAAccoert 15740	(Brief parameters of difference, threshold=0 corresponds to w Submit Queny Reset Results page TACCTTCTCTGCTGGCTCGGTGGGGGGAC GTCTCCGCGGGGGGCTTGGGCGGGGGGAGA	escription can be found on the <u>help page</u>) perfect matches)
3. Se 4. Sp Specific Or chi (read - Your Thress DNA c Your Your Quers Hits	lect DNA param ecify threshold v ly output format: Fas eck out our example our "how to start" by progress bar Ready shold=0.050000 coding scheme u pattern: SDRFSA stand acccAsGeccaccGecGer GGATCTCAGGCTTC y sequence is of found:	eter: 0 MeltEnthalpyBr rahe (0 ≤ Thr < 1) sta (default, recommend :: pre-selected pattern, p 25). sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGAGACT GGGAAGACTGGAAGAACC chr3, length=19950	eslauer 0.05 (normalized c or Table-vie parameters and <i>E. co</i> PBreslauer rGAGGAGTCGTCCTC rGGCTCAGAACCGCTT 15740	(Brief parameters of I difference, threshold=0 corresponds to w © th K12 genome as target sequence: © Submit Ouery Reset Results page TACCTTCTCTGCTGGCTGGGGGGAC GTCTCCGCGGGGGCTGGGGGGGACA	<pre></pre>
3. Se 4. Sp Specif Or chd (read.) Your I Threes DNA co CCGC// Your So CCGC// Quers No	lect DNA param ecify threshold v y output format: Fas eck out our example our "how to start" the progress bar Ready shold=0.050000 odding scheme u pattern: SDHRSA stand aGCCAAGGCCGACGG rogATCTCAGGCTTCC y sequence is found: Position (bp)	eter: 0 MeltEnthalpyBr ralue (0 ≤ Thr < 1) sta (default, recommend : pre-selected pattern, p sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGCCCCT CGGAAGACTGGAAGAACC chr3, length=19950 Similarity distar	eslauer [0.05] (normalized ed) C or Table-vie barameters and <i>E. co</i> "Breslauer "Breslauer "GGCTCAGAACCGCTT" 105740 102 [References	(Brief parameters of I difference, threshold=0 corresponds to w © <i>lit K12</i> genome as target sequence: Submit Query Reset Results page TACCTTCTTCTGCTGGCTGGGGGGGA GTCTCCCGCGGGGCTTGGGCGGGGGAGA Sequence ABAACACAAAAATTACCCGGGCTGGGGGGGGG	
3. Se. 4. Sp Specif Or chd (read.) Your Three Second Your	lect DNA param ecify threshold v y output format: Fast eck out our example our "how to start" tip progress bar Ready shold=0.050000 coding scheme us pathesh stand GCCAAGGCCGACGG rGGATCTCAGGCTTCA y sequence is found: Position (bp) 56230	eter: 0 MeltEnthalpyBr rahue (0 ≤ Thr < 1) sta (default, recommend :: pre-selected pattern, p sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGGGGCGCT CGGAAGACTGGAAGAACC chr3, length=19950 Similarity distar 0.04875	eslauer [0.05] (normalized c) C or Table-vie parameters and <i>E. co</i> /Breslauer rtGAGGAGTCGTCCTC GGCTCAGAACCGCTT 05740 Ince [References] Ensembl UCS	C (Brief parameters of d difference, threshold=0 corresponds to w C the K12 genome as target sequence: C Submit Query Reset Results page TACCTTCTCTGCTGGCTGGCGGGGAC GTCTCCGCGGGGGCTTGGGGGGGAC Sequence AAAAACACAAAAAATTACCCGGGCGTGGC GGGAACCGGGGGGGGGG	
3. Se 3. Se Specif Or chi (read Your I Three CCCCC Quers No 0 1	lect DNA param ecify threshold v y output format: Fas eck out our example our "how to start" by progress bar acady shold=0.050000 pattern: SDRFSA stand GGATCTCAGGCTTCC y sequence is found: Position (bp) 56230 56231	eter: 0 MeltEnthalpyBr rahue (0 ≤ Thr < 1) sta (default, recommend : pre-selected pattern, p sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGCGCGC chr3, length=1995C Similarity distar 0.04875 0.04680	eslauer [0.05] (normalized ed) C or Table-vie parameters and <i>E. co</i> pBreslauer rtGAGAGTOGTCCTC rGGCTCAGAACCGCTT J5740 [Ensemb1] UCS [Ensemb1] UCS [Ensemb1] UCS] [Ensemb1	(Brief parameters of (C) (C)	
3. Se 4. Sp Specif Cor chú (read.) Your Three Specif Quers No Quers No Quers No Quers No Quers No	lect DNA param ecify threshold v y output format: Fas eck out our example our "how to start" by progress bar Ready shold=0.050000 roding scheme us pattern: DDH5A stand GGATCTCAGGCTTCC y sequence is found: Position (bp) 56230 56231 56247	eter: 0 MeltEnthalpyBr rahue (0 ≤ Thr < 1) ts (default, recommend : pre-selected pattern, p sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGGCGCT CGGAAGACTGGAAGAACC chr3, length=1995C 5imilarity distar 0.04875 0.04680 0.04846	eslauer [0.05] (normalized ed) C or Table-vie parameters and <i>E. co</i> versauers and <i>E. co</i> versauers reaction of the second of the second reaction of the second of	Creating and a construction of the second seco	
3. Se 4. Sp Specif Or chi (read - ****F Threes DNA cr Scccccr Rule Query Hits 2 1 1 2 3	lect DNA param ecify threshold v y output format: Fast eck out our example our "how to start" tip progress bar Ready shold=0.050000 coding scheme us pattern: SHR5A stand AGCCAAGGCCGACGG FGGATCTCAGGCTTCC y sequence is found: Position (bp) 56230 56231 56248	eter: 0 MeltEnthalpyBr rahue (0 ≤ Thr < 1) sta (default, recommend : pre-selected pattern, p 25). sed - MeltEnthalpy ard; DNA; ROD; GTAGACGCTGGGGGGGCGCT CGGAAGACTGGAAGAACC chr3, length=1995C Similarity distar 0.04875 0.04886 0.04991	eslauer [0.05] (normalized or Table-vie parameters and <i>E. co</i> PBreslauer TGAGAGTCGTCCTC GGCTCAGACCGCTT 15740 [Ensembl] UCS [Ensembl] UCS [Ensembl] UCS	Create Construction Constr	

Figure 1. FeatureScan web interface (a) and sample output (b).

Parameter name	Туре	Unit	Comment						
MeltEnthalpyBreslauer normMeltEnthalpyBreslauer	Physico-chemical Physico-chemical	kcal/mol kcal/mol	Describes the melting of DNA double strands, values taken from (12) Normalized values of melting enthalpy						
Roll	Conformational angle	Degree	Describes respective bend angle of DNA (graphical view is given on the website)						
Twist	Conformational angle	Degree	Similar to roll						
Major groove depth Minor groove depth	Conformational linear dimension Conformational linear dimension	Å (Ångstrom) Å (Ångstrom)	Describes the depth of the major groove of DNA Describes the depth of the minor groove of DNA						

Table 1. Description of a few DNA parameters, including those which were used in Refs (10-12) and which describe the three major groups of DNA features

1 vs. 2 Letter identity: 44/97 (45.4%)

Signal identity: 0.88

1 vs. 3 Letter identity: 5	57/109 ((52.3%)
----------------------------	----------	---------

Signal identity: 0.85

		•		•									-							
TAGATAT	CTAAAAGGAT	TTAGC	CAACTAT	C-TT	TTT	TAT	TTAT	TGA	AAAG	GATAAA	GCAT	ГАААА	CTG	GGA		-AAC	TATA	AAGTA	4ATG7	ATATA
TCAGTCC	AGATAAAATG	ATGTC	CAACGAT	TGTT	TCT	GATI	TTAT	TAGA	CAAA	TAAAA	CCGT	ГАААА	CAGI	rgagt	TTTACT	CAAT	TAAT	GAGAG	TTTG	CTATG

Figure 2. Similarity revealed by FeatureScan and BLAST of promoters of *E.coli* genes belonging to class 6.3 'Surface antigens' (according to GenProtEC database (7): (i) promoter region of UDP-D-glucose:(glacosyl)lipopolysaccharide glucosyltransferase gene; (ii) promoter region of UDP-D-galactose:(glucosyl)lipopolysaccharide-alpha-1,3-D-galactosyltransferase gene; and (iii) promoter region of UDP-D-galactose:(glucosyl)lipopolysaccharide-1, 6-D-galactosyltransferase gene. (ii) versus (iii): Letter identity, 0.52; signal identity, 0.89; alignment not shown.

reflecting respective changes in physical properties behind them.

A fully annotated *E.coli* genome was chosen to investigate the ability of our technique to reveal functional similarities between promoter regions (11). We built up a similarity matrix of *E.coli* promoters. Subsequent statistical analysis of the detected promoter similarities and of the cellular functions using the GenProtEC E.coli database (14) identified an enrichment of several functional categories of genes, the promoters of which showed high signal similarity. It was interesting to observe that among the functional classes most populated by similar promoters were classes of genes which encode membrane-specific proteins such as surface antigens and transporters. Cross checked with BLAST, these promoters showed similarity only at a background level (see typical example in Figure 2). One may speculate that the conformational parameter 'roll', which was used in this analysis, best reflects the requirements for specific 3D structure of the promoters for effective binding and assembling of the transcription machinery.

We performed a comparison of two evolutionary close genomes—human and chimpanzee. Similarity between promoters of homologous genes was calculated by ClustalW and FeatureScan (the latter by using DNA property melting enthalpy). We found that homologous promoters are much closer by their melting characteristics than by the number of conserved nucleotides (I. V. Deyneko, A. E. Kel, J. M. Kalybaeva, H. Blöcker and G. Kauer, unpublished data). This observation may indicate that mismatches in promoters were introduced in a correlated manner during the course of evolution, so that the DNA property 'melting enthalpy' of the promoters is retained.

Such promising advantages of property-dependant similarity measures encourage us to look for further application. For example, SNP mutations which are seen by Blast-type methods as single nucleotide mismatches, may introduce a number of different changes to DNA characteristics (melting temperature, conformation and others). FeatureScan is designed and can specifically spot such alterations and thus provide new insights into biological functionalities.

CONCLUSION

Here we present web resources for sequence analysis based on a novel type of sequence similarity. We show that it has certain specific advantages over letter-based methods. We think that none of the alternate approaches can entirely replace the other and that both will evolve in parallel. FeatureScan can be easily applied to the analysis of other information-bearing biomolecules by just using an appropriate (physical nature reflective) letter-to-signal encoding scheme. In the future, methods utilizing signal-theoretical similarity may be further automated and trained specifically for recognition of promoters, binding sites and others. It should be noted that utilizing specific hardware enables one to carry out search procedures with huge amounts of eukaryotic data in acceptable time.

ACKNOWLEDGEMENTS

This project is part of the Bioinformatics Competence Centre 'Intergenomics'. Generous financial support by the German Federal Ministry for Education and Research through Projektträger Jülich (FKZ 031U210A) and through the 'Impuls- und Vernetzungsfond' of the Helmholtz Association (FKZ VH-VI-023) is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by GBF.

Conflict of interest statement. None declared.

REFERENCES

- Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
- Chekmenev,D.S., Haid,C. and Kel,A.E. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.*, 33, W432–W437.
- Shahmuradov,I.A., Solovyev,V.V. and Gammerman,A.J. (2005) Plant promoter prediction with confidence estimation. *Nucleic Acids Res.*, 33, 1069–1076.
- Gangal, R. and Sharma, P. (2005) Human pol II promoter prediction: time series descriptors and machine learning. *Nucleic Acids Res.*, 33, 1332–1336.
- 5. Marshall,J.L. (1965) *Introduction to Signal Theory*. International Textbook Co., Scranton, PA.
- Ponomarenko, J.V., Ponomarenko, M.P., Frolov, A.S., Vorobyev, D.G., Overton, G.C. and Kolchanov, N.A. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, 15, 654–668.
- Kauer,G. and Blöcker,H. (2004) Analysis of disturbed images. Bioinformatics, 20, 1381–1387.

- 8. Rabiner,L. and Juang,B.H. (1993) *Fundamentals of Speech Recognition*. Prentice Hall, NJ.
- Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1998) *The Art of Scientific Computing*. Cambridge University Press, Cambridge.
- Kauer,G. and Blöcker,H. (2003) Applying signal theory to the analysis of biomolecules. *Bioinformatics*, 19, 2016–2021.
- 11. Deyneko, I.V., Kel, A.E., Blöcker, H. and Kauer, G. (2005) Signaltheoretical DNA similarity measure revealing unexpected similarities of *E.coli* promoters. *In Silico Biol.*, **5**, 0049.
- Breslauer,K.J., Frank,R., Blöcker,H. and Marky,L.A. (1996) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci.* USA, 83, 3746–3750.
- Deyneko,I.V., Kel,A.E., Wingender,E., Gössling,F., Blöcker,H. and Kauer,G. (2004) Signal theory—an alternative perspective of pattern similarity search. In *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, Russia, vol. 2, pp. 25–28.
- Serres, M.H. and Riley, M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comput. Genomics*, 5, 205–222.