SAR and QSAR in Environmental Research,
Vol. 14 (5–6), October–December 2003, pp. 413–420

Taylor & Francis
Taylor & Francis Group

A NOVEL METHOD FOR GPCR RECOGNITION AND FAMILY CLASSIFICATION FROM SEQUENCE ALONE USING SIGNATURES DERIVED FROM PROFILE HIDDEN MARKOV MODELS*

P.K. PAPASAIKAS, P.G. BAGOS, Z.I. LITOU and S.J. HAMODRAKAS†

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece

(Received 13 July 2003; In final form 30 August 2003)

G-protein coupled receptors (GPCRs) constitute a broad class of cell-surface receptors, including several functionally distinct families, that play a key role in cellular signalling and regulation of basic physiological processes. GPCRs are the focus of a significant amount of current pharmaceutical research since they interact with more than 50% of prescription drugs, whereas they still comprise the best potential targets for drug design. Taking into account the excess of data derived by genome sequencing projects, the use of computational tools for automated characterization of novel GPCRs is imperative. Typical computational strategies for identifying and classifying GPCRs involve sequence similarity searches (e.g. BLAST) coupled with pattern database analysis (e.g. PROSITE, BLOCKS). The diagnostic method presented here is based on a probabilistic approach that exploits highly discriminative profile Hidden Markov Models, excised from low entropy regions of multiple sequence alignments, to derive potent family signatures. For a given query, a *P*-value is obtained, combining individual hits derived from the same family. Hence a best-guess family membership is depicted, allowing GPCRs' classification at a family level, solely using primary structure information. A web-based version of the application is freely available at URL: http://bioinformatics.biol.uoa.gr/PRED-GPCR.

Keywords: GPCR; Family; Prediction; Classification; Signatures; Profile HMMS

INTRODUCTION

G protein coupled receptors (GPCRs) constitute a vast cell surface receptor family, populated with hundreds of members with versatile functions. GPCRs are also referred to as seven-transmembrane receptors, because of their characteristic configuration of an anticlockwise bundle of 7 transmembrane α helices, signal by activating heterotrimeric G proteins, although alternative signalling pathways have recently been described [1]. The upstream position of GPCRs in various basic regulatory pathways along with their accessibility, due to their membrane localization, have designated GPCRs as a highly amenable target class to therapeutic intervention. Furthermore, the completion of various genome and proteome

*Presented at CMTPI 2003: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Thessaloniki, Greece, September 17–19, 2003).

†Corresponding author. E-mail: shamodr@cc.uoa.gr

ISSN 1062-936X print/ISSN 1029-046X online
DOI: 10.1080/1062936031000162399 9

projects is proving to be a driving force for the growing role of GPCRs in drug discovery. These projects have yielded a drove of putative GPCRs with unknown functions and ligands, better known as "orphan" GPCRs. In the absence of experimental data, computational methods are the simplest and most frequently used strategy when trying to identify and attain functional specificity to novel receptors.

GPCRs can be grouped into six distinct classes A,B,C,D,E and frizzled/Smoothened family as defined by GPCRDB classification on the base of shared sequence motifs [2]. Traditionally, though, GPCRs' family classification is based on receptor's ligand specificity. However, common ligand specificity does not necessarily infer a certain level of sequence identity. Somehow some higher-order relationship between sequence and binding of ligands of a particular chemical class seems to exist, as revealed by phylogenetic analysis [3]. Nevertheless, no clear correlation between sequence similarity and ligand specificity can be assessed; sequences within the same family are often heterogeneous, sharing as little as 25% identity to one another (e.g. members of the histamine receptor family), whereas others, with a high degree of overall similarity can have little in common when it comes to function. To tangle things more, some GPCRs appear to be, partly, mosaics of modular domains, repeated among families functionally distinct [4]. These observations reflect a complex evolutionary background with GPCR sequences converging, diverging or following parallel branches, before they come up with their current functional profile.

Database search methods based on pairwise similarity (e.g. BLAST [5]) do not take such subtleties into account since they only appreciate generic similarities between sequences. As a result, top-scoring matches cannot safely be assumed functional analogues of a query sequence.

Pattern databases such as PROSITE [6], Pfam [7], BLOCKS [8], PRINTS [9], and InterPro [10] were built in order to overcome such problems. The common approach behind these databases is to deploy key regions, as "baits" in order to distinguish between families, instead of using the whole sequence. These regions, often referred to as signatures, characterize one family and highly improve the diagnostic performance of automated classification. Moreover, signatures often indicate functional or structural determinants within a certain family.

Several methods are used to encode the information of such regions. These methods include regular expressions (PROSITE [6]), position-specific scoring matrices (BLOCKS [8]), frequency matrices (PRINTS [9]) and profile hidden Markov models (Pfam [7]).

Profile hidden Markov models (profile HMMs) are statistical models of the primary structure consensus of a sequence family [11]. From the methods mentioned above, profile HMMs is the only one that has a formal probabilistic basis and the only method for encoding gapped regions of alignments, which could also contain crucial information [11,12]. Karchin *et al.* [13] showed that profile HMMs can accurately distinguish among GPCR superfamilies.

Highly informative profile HMMs may be constructed either from full-length multiple sequence alignments (MSAs) or conserved MSA segments. However, such profiles turned out not to perform particularly well when applied to higher-level GPCRs-classification problems. This should have been expected, considering that:

a. higher-level GPCR-classification is not based solely on sequence similarity and
b. many GPCRs with analogous functions have resulted from convergent evolution.

This is why only a few profiles that correspond to GPCR ligand specificity are currently available in Pfam.

The approach proposed in our method exploits the descriptive power of profile HMMs along with an exhaustive discrimination assessment method to select only highly selective

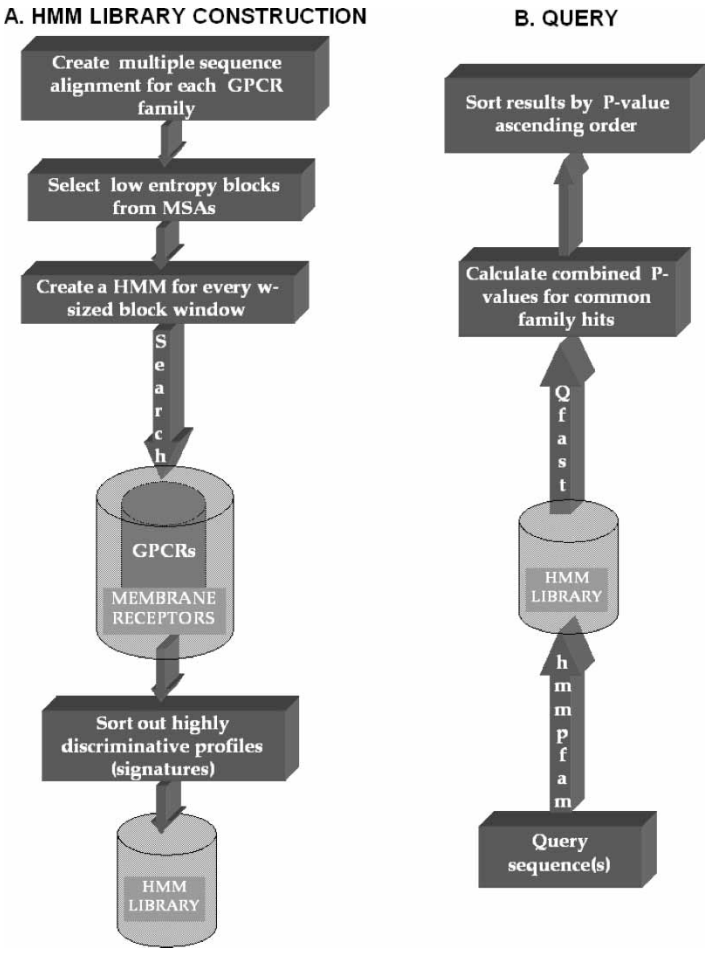

FIGURE 1 Flow diagram of the method. (A) HMM library construction. First an MSA is constructed for each family subset. Then MSA blocks of low entropy are selected and all overlapping windows with a predefined maximum width are created. For each window a HMM is created. The performance of all HMMs as family classifiers is estimated after a *hmmsearch*- pass through the membrane receptors data set. Only those HMMs with high selectivity and sensitivity for the family are selected for the HMM library. (B) Query. A given sequence or set of sequences is queried against the HMM library using the *hmmpfam* program. Combined *P*-values for each family are obtained, implementing the Qfast algorithm. The results are reported ranked with a *P*-value ascending order.

and sensitive profiles, after mass-constructing confined conserved blocks of MSAs for each family. The collection of these profiles constitutes a signature library, which is scanned in order to determine to which GPCR family a query sequence belongs or resembles (Fig. 1).

MATERIALS AND METHODS

Primary Data

In this work we used a set of 1866 well-annotated eucaryotic receptors, excluding fragments, carefully collected from the Swiss-Prot and TrEMBL databases [14]. Of these, 965 were GPCRs with known ligands, partitioned into subsets of 65 families based on the pharmacological classification of receptors [15] and the GPCRDB information system [2]. These subsets were used for the construction of each family's multiple sequence alignment.

The remaining 901 sequences were non-GPCR membrane receptors used, along with the 965 GPCR sequences, for the evaluation of each profile HMM as a classifier of a particular GPCR family.

HMM Library Construction

Selecting Low Entropy Blocks from Multiple Alignments

Initially, multiple sequence alignments were constructed for each GPCR family, using the ClustalX 1.81 package [16]. Pairwise alignment parameters were set as: an opening gap penalty of 10, an extension gap penalty of 0.1 and a Blosum 30 scoring matrix [17]. Multiple alignment parameters were set as: an opening gap penalty of 10, an extension gap penalty of 0.2 and the Blosum series of scoring matrices [17].

ClustalX provides an indication of the quality of an alignment by assigning a normalized "conservation score" (S_c) for each column of the alignment, which varies from 0 to 100 [16] (Fig. 2).

Low entropy "cores" of 5 contiguous columns with $\bar{S_c} \geq \mathrm{Th_c}$ were detected. The threshold $\mathrm{Th_c}$ was empirically set to 30 as below this value performance did not seem to improve. For these "cores" both flanking regions are examined and extended until they are surrounded by at least 3 consecutive non-conserved columns at both flanking regions. A column is considered as non-conserved if $S_c \leq 20$. This value was set lower than $\mathrm{Th_c}$ to ensure that no critical columns, near a conserved MSA fragment, will be lost during block construction. This procedure excludes divergent segments where alignment is either ambiguous or saturated by multiple substitutions [18,19] resulting in low entropy blocks of varying length. For each conserved block, sliding MSA windows with a predefined maximum width of 21 columns were created (Fig. 2). For each MSA window an HMM was built and calibrated with the HMMER (v 2.2) software package [20]. A relatively small value of length for

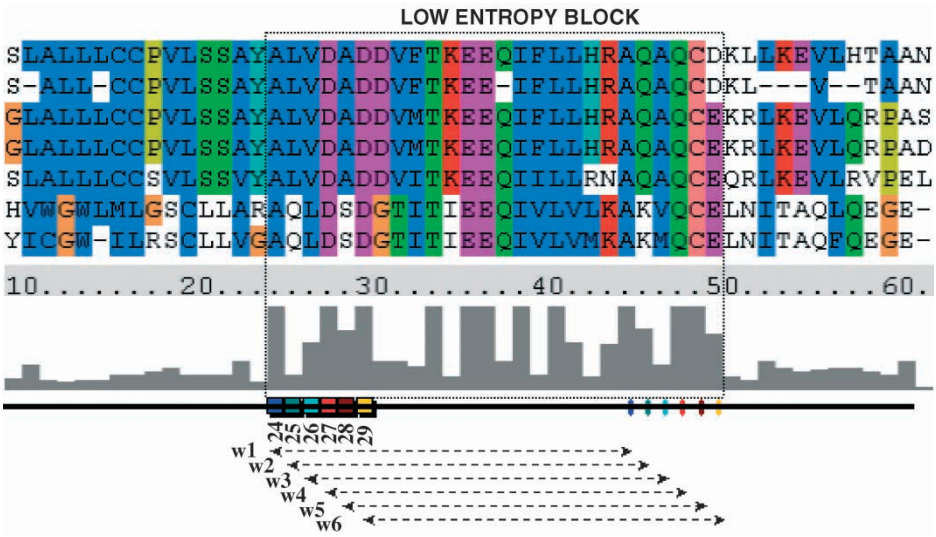

FIGURE 2 A small fragment of the MSA of the parathyroid hormone receptor family built with ClustalX. Below the MSA, ClustalX provides a histogram for the columns' conservation scores. Blocks of low entropy from the MSAs are selected with a computerized method that excludes positions where the alignment is ambiguous or is likely to be saturated with multiple substitutions. All possible MSA windows with a predefined maximum width of 21 columns are created for each block. The start and end positions (colored squares and bars respectively) for the six possible MSA windows of this block are depicted below the histogram.

the HMMs is desirable in order to bring out and capture subtle differences between closely related families, lying in narrow key-regions of the sequence [21]. The forfeit is that the informative content and, therefore, the absolute scoring potency of such models is confined. However the use of multiple signatures for each family can compensate for the low informative content of each particular profile.

Evaluating the Performance of Profile HMMs as Family Classifiers

The performance of each HMM as a classifier of a family with n members was estimated after an *hmmsearch*-pass through the membrane receptors data set. Therefore, for each family this data set consists of n positive examples and 1866-n negative examples (see Primary data). The *hmmsearch* program of the HMMER package returns a list of all matching sequences in ascending order of expectation values (E-values) [20]. The threshold E-value was arbitrarily set to 1 and selection of those HMMs to join the family specific HMM signatures library was based upon the following criteria, assessed in a hierarchical manner:

1. The calculated value of an empirical estimator of discriminative performance, DE. Albeit all true positives (TP) are considered equally important, early raises of the false positive rate (FPR, FP/negatives), are much more harmful when trying to spotlight family signatures. Therefore, we used coverage (percentage of true positives, before the first false positive error) to measure sensitivity and a logarithmic transformation of FPs, normalized for all negatives, to measure specificity. As a result the estimator is unbiased for sensitivity and strongly weighted for high specificity:

$$\mathrm{DE} = \mathrm{C} - \frac{\log(\mathrm{FP}+1)}{\log(N-n+1)},$$

   where N is the size of the data set (1866), n the number of family members, C is the coverage (percentage of true positives, before the first false positive error) and FP is the number of false positives. Note that the value $\log(N-n+1)$ is a constant for each family. The fractional term of the function DE rises sharply for the first few FPs, whereas it becomes insensitive for extreme values of FP. Since both terms of the function (coverage, C and the fractional term) are regularized, DE varies from -1 to 1.
2. The E-value of the first False positive hit (FFP).
3. The E-value of the last True positive hit (LTP).

In the case of profiles derived from overlapping segments of the MSA, only the best one was selected, according to the criteria discussed above, in order to ensure P-value independence for a given query. This is important if a combined P-value for all common family hits is to be calculated. P-value measures consistency between the results actually obtained in the trial and the "pure chance" explanation for those results. The P-value of a match of a sequence to a motif is defined as the probability, of a randomly generated sequence of the same length, having a match score at least as high as the match score of the sequence.

Querying the HMM Library

For a given unknown sequence, or set of sequences, the query is performed against the HMM library, using the *hmmpfam* program of the HMMER software package [20]. *hmmpfam* reads a sequence file and compares each sequence in it, one at a time, against all the HMMs in the library looking for significantly similar sequence matches [20]. For all significant hits

pointing to the same family a combined *P*-value is obtained, implementing the Q-fast algorithm [22]. Q-fast provides a statistically valid method for combining sources of evidence with independent, continuous distributions. A combined *E*-value for each family can then be calculated, assuming a database size of $\langle m \rangle$ sequences. Thus a measure of the statistical significance, of a match of a sequence to all profiles derived from the same family, is provided.

RESULTS AND DISCUSSION

To date the library houses a collection of more than 200 profile-HMMs for 65 families of GPCRs. In order to confirm the specificity of our method we tested it on a set, which consists of 1239 globular and 1361 non-GPCR transmembrane proteins with less than 25% pairwise similarity. The combined *E*-value threshold was set to 0.03, which was the minimum error point (MEP) for the membrane receptors data set. MEP is the *E*-value threshold where classifiers make the fewest errors (false positives plus false negatives). Our method misclassified only 0.4% of these negative examples. The results for this data set are summarized in Table I.

To demonstrate the efficacy of the method we have applied it to an independent set of 310 well-annotated sequences of GPCRs, recently deposited in the Swiss-Prot database, excluding fragments. These sequences were not included in the primary data set. Using the same *E*-value threshold, our method classified correctly 298 of these sequences (success rate 96%).

An online application of this method is freely accessible at URL: http://bioinformatics.biol.uoa.gr/PRED-GPCR. Available options for this application include selection of an *E*-value threshold and a filter that implements the CAST algorithm [23], allowing, low-complexity region detection and selective masking. This filter can significantly improve the selectivity of the method, since the sequence score takes into account all scoring domains and could, therefore, return false positives in case of low scoring domains repeated along the sequence.

An example of the search output for two query sequences is shown in Fig. 3. A separate report is returned for each input sequence. This report consists of two sections: a ranked list of the profile HMM matches, below the selected *E*-value cut-off, along with their corresponding family, and a ranked list of the combined *P*-values, *E*-values and the number of profiles matched for each family (Fig. 3). For the combined *E*-value calculation we assume the HMMER default database size (59021) [20].

These results demonstrate the discriminative potency of this method and its probable use as a screening tool for dissecting new members of known GPCR families from recently sequenced genomes.

TABLE I Family recognition results on a test set of 1361 non-GPCR transmembrane proteins and 1239 globular proteins with less than 25% pairwise similarity

	Prediction		
	Number of unassigned sequences	*Number of falsely assigned sequences*	*Incorrect assignment %*
TM.	1353	8	0.6%
GLOB.	1236	3	0.2%
TOTAL	2589	11	0.4%

The combined *E*-value threshold was set to 0.03, which was the MEP (see text) for the primary data set.

sw|P28678|OPS1_DROPS

Profile	Family	E-value
op51	Opsin	3.4e-10
op101	Opsin	6.6e-08
op21	Opsin	4.3e-06
op83	Opsin	0.081
ad136	Adenosine rec	0.17

FAMILY DESCRIPTION	Combined p-value	Combined e-value	Family profiles
Opsin	1.02e-37	5.99e-33	4 out of 4
Adenosine receptor	2.88e-06	1.70e-01	1 out of 3

sw|P08100|OPSD_HUMAN

Profile	Family	E-value
op51	Opsin	2.2e-10
op101	Opsin	8e-09
op21	Opsin	2.2e-07
op83	Opsin	6.7e-06

FAMILY DESCRIPTION	Combined p-value	Combined e-value	Family profiles
Opsin	5.15e-44	3.04e-39	4 out of 4

FIGURE 3 Search report for two query sequences. For each query sequence the report consists of two sections: a ranked list of all profile HMM matches, below the selected *E*-value threshold, along with their corresponding family, and a ranked list of the combined *P*-values, *E*-values and the number of profiles matched for each family.

A very important attribute of this method is that no structural or functional information is imposed *a priori* during the selection process of signature profile HMMs. Such information is used, for example, for the Prints database GPCR-signatures selection [9]. Therefore, following a reverse rationale, this method could offer some insight beyond the primary structure level; further research is needed to explore the probable role of these sequence segments, from which such discriminative profiles were derived, in determining the unique structural and functional profile of each GPCR-family.

References

[1] Pierce, K.L., Premont, R.T. and Lefkowitz, R.J. (2002) "Seven-transmembrane receptors", *Nat. Rev. Mol. Cell Biol.* **9**, 639–650.

[2] Horn, F., Weare, J., Beukers, M.W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F. and Vriend, G. (1998) "GPCRDB: an information system for G protein-coupled receptors", *Nucleic Acids Res.* **26**, 275–279 [http://www.gpcr.org/7tm].

[3] Vassilatis, D.K., Hohmann, J.G., Zeng, H., Li, F., Ranchalis, J.E., Mortrud, M.T., Brown, A., Rodriguez, S.S., Weller, J.R., Wright, A.C., Bergmann, J.E. and Gaitanaris, G.A. (2003) "The G protein-coupled receptor repertoires of human and mouse", *Proc. Natl Acad. Sci. USA* **100**, 4903–4908.

[4] Stacey, M., Lin, H.H., Gordon, S. and McKnight, A.J. (2000) "LNB-TM7, a group of seven-transmembrane proteins related to family-B G-protein-coupled receptors", *Trends Biochem. Sci.* **25**, 284–289.

[5] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* **25**, 3389–3402.

[6] Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. and Bairoch, A. (2002) "The PROSITE database, its status in 2002", *Nucleic Acids Res.* **30**, 235–238.
[7] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) "The Pfam protein families database", *Nucleic Acids Res.* **30**, 276–280, 2002.
[8] Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. (2000) "Increased coverage of protein families with the blocks database servers", *Nucleic Acids Res.* **28**, 228–230.
[9] Attwood, T.K., Croning, M.D. and Gaulton, A. (2002) "Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors", *Protein Eng.* **15**, 7–12.
[10] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R. and Zdobnov, E.M. (2003) "The InterPro Database, 2003 brings increased coverage and new features", *Nucleic Acids Res.* **31**, 315–318, 2003.
[11] Eddy, S.R. (1998) "Profile hidden Markov models", *Bioinformatics* **14**, 755–763.
[12] Hughey, R. and Krogh, A. (1996) "Hidden Markov models for sequence analysis: extension and analysis of the basic method", *Comput. Appl. Biosci.* **12**, 95–107.
[13] Karchin, R., Karplus, K. and Haussler, D. (2002) "Classifying G-protein coupled receptors with support vector machines", *Bioinformatics* **18**, 147–159.
[14] Bairoch, A. and Apweiler, R. (2000) "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", *Nucleic Acids Res.* **28**, 45–48.
[15] Alexander, S.P.H. and Peters, J.A. (2000) TiPs Receptor and Ion Channel Nomenclature Supplement, Vol. **11** (Elsevier).
[16] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools", *Nucleic Acids Res.* **25**, 4876–4882.
[17] Henikoff, S. and Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks", *Proc. Natl Acad. Sci. USA* **89**, 10915–10919.
[18] Goldman, N. (1998) "Phylogenetic information and experimental design in molecular systematics", *Proc. R. Soc. Lond. B. Biol. Sci.* **265**, 1779–1786.
[19] Yang, Z. (1998) "On the best evolutionary rate for phylogenetic analysis", *Syst. Biol.* **47**, 125–133.
[20] Eddy, S.R. (2000) HMMER: profile hidden Markov models for biological sequence analysis (Washington University school of medicine, St Louis, MO) [http://hmmer.wustl.edu].
[21] Truong, K. and Ikura, M. (2002) "Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach", *BMC Bioinformatics* **3**, 1.
[22] Bailey, T.L. and Gribskov, M. (1998) "Combining evidence using p-values: application to sequence homology searches", *Bioinformatics* **14**, 48–54.
[23] Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C. and Ouzounis, C.A. (2000) "CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts", *Bioinformatics* **16**, 915–922.