BIOINFORMATICS

Vol. 16 no. 10 2000
Pages 915–922

CAST: an iterative algorithm for the complexity analysis of sequence tracts

Vasilis J. Promponas[1], Anton J. Enright[2], Sophia Tsoka[2], David P. Kreil[3], Christophe Leroy[4], Stavros Hamodrakas[1], Chris Sander[4] and Christos A. Ouzounis[2,*]

[1]Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens GR-15701, Greece, [2]Computational Genomics Group, [3]SRS Team, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK and [4]Millennium Pharmaceuticals Inc., 640 Memorial Drive, Cambridge, MA 02139, USA

Received on February 8, 2000; revised on April 11, 2000; accepted on April 18, 2000

Abstract

Motivation: *Sensitive detection and masking of low-complexity regions in protein sequences. Filtered sequences can be used in sequence comparison without the risk of matching compositionally biased regions. The main advantage of the method over similar approaches is the selective masking of single residue types without affecting other, possibly important, regions.*

Results: *A novel algorithm for low-complexity region detection and selective masking. The algorithm is based on multiple-pass Smith–Waterman comparison of the query sequence against twenty homopolymers with infinite gap penalties. The output of the algorithm is both the masked query sequence for further analysis, e.g. database searches, as well as the regions of low complexity. The detection of low-complexity regions is highly specific for single residue types. It is shown that this approach is sufficient for masking database query sequences without generating false positives. The algorithm is benchmarked against widely available algorithms using the 210 genes of* Plasmodium falciparum *chromosome 2, a dataset known to contain a large number of low-complexity regions.*

Availability: *CAST (version 1.0) executable binaries are available to academic users free of charge under license. Web site entry point, server and additional material: http://www.ebi.ac.uk/research/cgg/services/cast/*

Contact: *ouzounis@ebi.ac.uk*

Introduction

The explosion of sequence information requires extensive sequence comparison for the detection of homologies and the prediction of function using sequence similarity. One of the most widely used approaches and a necessary step for any further analysis is the searching of databases using newly sequenced proteins as queries for the identification of homologues.

Rapid and sensitive algorithms have been developed to perform homology searches in sequence databases, such as BLAST (Altschul *et al.*, 1997) or FASTA (Pearson, 1990). When high sequence similarity is detected between a query sequence and a well-characterized database entry, a reliable function prediction for the query sequence can be obtained.

However, some of the high-scoring database entries may contain compositionally biased regions of amino acid residues, also known as 'low-complexity' regions (Wootton, 1994). These hits may result in erroneous function predictions, because the sequence similarity is due to this effect and not necessarily to genuine homology. The terms 'compositionally biased' and 'low-complexity' regions are interchangeably used herein.

Algorithms that filter sequences for low-complexity regions have been developed to effectively process queries before a sequence database search is performed. Two such methods are XNU (Claverie and States, 1993) and SEG (Wootton and Federhen, 1993). XNU identifies repeats on the basis of a self-comparison of the query sequence (Claverie and States, 1993), while SEG detects low-complexity regions based on an information measure (Wootton and Federhen, 1993). Both methods alter the query sequence by replacing the low-complexity regions with X symbols (undefined residue type) as whole segments, a process known as 'masking'. The masking of biased regions has significantly improved the reliability of homology detection and, consequently, the quality of function predictions by homology.

Here we describe a different method for low-complexity region detection and masking, called CAST (complexity

*To whom correspondence should be addressed.

analysis of sequence tracts). The method differs from previous approaches in that single residue types are detected and masked for database searching. As a by-product of the detection step, low-complexity regions are extracted with high specificity for single residue types and can be used for further study.

System and methods

The CAST program is written in ANCI C, and developed on a Sun Ultra 10 Workstation. The code has been ported to various platforms, including Solaris, SGI IRIX, AIX and Linux. Executables for other platforms are available from the authors on request. Compiled versions allow users to vary the algorithm parameters. A web server has also been implemented and is available via the WWW, starting at http://www.ebi.ac.uk/research/cgg/services/cast/.

Algorithm

Formulation of the problem

In contrast to previous work, low-complexity regions in protein sequences are defined by an empirical criterion, as those regions that score high in homology searches with degenerate sequences composed of a single amino acid type (homopolypeptides). This formulation arises from general experience in database searches, where even the most degenerate sequences can give high scores to biased regions in absence of any real sequence similarity (Figure 1). This definition is reminiscent of previously described (but not further elaborated) approaches: in an extensive search for unidentified bacterial open reading frames, a collection of known compositionally biased proteins was used to detect sequences of unusual bias and exclude them (Robison *et al.*, 1994).

In the extreme formulation of this approach, homopolymeric peptides can be used as the baseline for the detection of low-complexity regions. By definition, homopolypeptides do not contain any real sequence information at all. Each one can be completely characterized by two values: its monomer type and length. Evidently, such a homopolymer will produce high similarity scores with proteins (or domains) of similar composition, not depending on the actual sequence of these proteins.

Supposing that the fractions of different residue types a, b in a search and a test sequence respectively are statistically unrelated events, then the probability to find a match of residue types a and b could be readily calculated from the independent residue frequencies as:

$$p_{ab} = f_a p_b \tag{1}$$

where f_a, p_b are the fractions of amino acid types a, b in the search and test sequences respectively.

Scoring all possible $a - b$ matches with a proper comparison matrix M composed of elements $m_{a,b}$, the average expected score over a region of length l would be:

$$l \sum_{a,b} (p_{ab} m_{a,b}) = l \sum_{a,b} (f_a p_b m_{a,b}). \tag{2}$$

Higher scores reflect similar sequence patterns. As we admit any local region of length l in both proteins, we can ignore the factor l. Consequently, the frequencies f_a and p_b correspond to the local residue frequencies in the two compared regions of the search (test) protein. If we consider a particular region in the test protein, the residue composition is invariant and the sum score over all residue types can be performed. Therefore, the only remaining variable is the composition of the search sequence $f_{a\beta}$ and equation (2) can be written as:

$$\sum_{a,b} (f_a p_b m_{a,b}) = \sum_a f_a \sum_b (p_b m_{a,b}) = \sum_a (f_a C_a) \tag{3}$$

where C_a is a parameter clearly related only to the residue type a in the search sequence.

Residue frequencies f_a are bounded between 0 and 1 and sum up to 1 (as required for random variables). Therefore, the last part of equation (3) is an interpolation between the 20 possible values C_a and can only result in scores between the smallest and the largest value of C_a. The maximum obtainable sum is the case where the sum is equal to the largest C_a, arbitrarily sorted as C_1. A general case, where the maximum score is always obtained, is when the corresponding residue frequency f_1 equals to 1, which corresponds to the homopolymer. Therefore, one of the 20 homopolymers will always have the highest score obtainable by any unrelated sequence.

Certainly, this argument does not apply if the two complex sequences share more than just a similar composition. In this case, the complex sequence can have much higher scores that reflect real sequence similarity. There is a well-developed statistical theory allowing the estimation of likelihood for such similarities to arise by chance (Karlin and Altschul, 1990).

Detection of low-complexity regions

Based on the above idea, the problem can be stated as searching an artificial database consisting of 20 degenerate protein sequences of arbitrary length, each one being a homopolymer based on one of the 20 natural amino acid residue types. A homology search of a protein against this 'bias-database' will only report significant hits, if the search protein contains regions of 'unusual' amino acid composition. The 'homologue' found in this search immediately identifies the type of bias and the 'region of homology' identifies the region of bias.

In analogy to this concept of a bias-database that can be used with any homology search program, the

```
Length = 810
Score = 91.3 bits (223), Expect = 2e-18
Identities=53/137(38%),Positives=70/137(50%),Gaps=17/137 (12%)

Q: 1   RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR 60
       +RRR   R  RRRR R R R R R RR    R  +R +  R+RR + RRRR + +R R R
S: 132 QRRREHEREERRRRERERERERGRGRRDENERDPKREQEERQRREQERRRREQEQRERER 191

Q: 61  RRRR-------------RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR----RR 103
       R  R             RR + RR + RRRR + + RR R+RR  R      +R    RR
S: 192 RGERDEEDDENQRDPDWRREQERREQERRRREQEQERRERQRRGGRDDEDENQRDPDWRR 251

Q: 104 RRRRRRRRRRRRRRRR 120  'Q: QUERY SEQUENCE'  [polyR polymer]
        ++RR + + RR R RR      IDENTITY: R; SIMILARITY: +
S: 252 EQKRREQEQERRERERR 268  'S: DATABASE TARGET' [gi|3808061]
```

Fig. 1. Example of comparing a homopolymer sequence against a natural protein sequence. BlastP output (BlastP version 2.0.8, Blosum62 Mutation Matrix) for a 120 residues long query 'poly-R' sequence against nrdb. More than 80 hits were reported with an E-value less than 10^{-11}. All Arginine-rich protein sequences score high with the artificial homopolymeric query sequence. The filtering option of Blast was turned off.

search against the 20 amino acid type homopolymers can be formulated as an algorithm using a dynamic programming procedure with the same steps used by most homology search programs. In this special case, the search procedure based on the Smith–Waterman algorithm (Smith and Waterman, 1981) can be simplified, because in homopolymers the position in the sequence is irrelevant. Therefore, only a single linear pass over the sequence is necessary for each residue type a in order to find regions of maximal score.

Assume a biological sequence $r_1, r_2, r_3, \ldots, r_i, \ldots, r_n$ of length n. A score at position $i (1 \leq i \leq n)$ is derived by adding the value counted at the previous position and the score given in the mutation matrix for a match of residue type a with the amino acid type r in the sequence at position i.

$$s_i^{\alpha} = m_{a,r_i} + \begin{cases} s_{i-1}^{\alpha}, & s_{i-1}^{\alpha} \geq 0 \\ 0, & s_{i-1}^{\alpha} < 0 \end{cases}. \qquad (4)$$

Regions of maximal score are identified as stretches in the resulting run of values that have positive scores and range from the first positive score to its maximum. Applying this algorithm for all residue types a, all 20 possible types of bias are detected. The score at the maximum gives a quantitative measure for the bias detected with a particular scoring matrix and the associated type of bias.

In this implementation, gaps are not allowed. This feature would only introduce more complexity to the algorithm by the choice of external parameters (the gap penalties) leading to arbitrary dependencies. According to the above formulation, the algorithm is equivalent to multiple passes of a Smith–Waterman comparison between the query sequence and a set of homopolymers of equal length, with infinite gap open and elongation penalties.

Masking low-complexity regions using CAST

To eliminate high scoring matches caused by composition bias, sequences are prepared for homology searches by masking biased regions. With currently available methods (Claverie and States, 1993; Wootton and Federhen, 1993), masking is done by completely replacing detected biased regions with stretches of undefined residue type X (Figure 2). These symbols are treated as 'unknowns' in sequence comparison. Therefore, matches in this region to other sequences are drastically down-weighted and non-biased regions of the proteins become more prominent in database searches.

The CAST algorithm not only detects regions of composition bias but also identifies the type of residue a causing the bias. With this additional information, masking can be done in a much more subtle and specific manner. Only the residue type over-represented in the biased region is replaced by undefined residue type X, hereafter called '*bias type*'. All other residues remain unmodified. Therefore, part of the sequence information is saved even after correcting for the composition bias and can still be matched to homologous proteins, as illustrated in Figure 2.

As this procedure permits the use of a comparison matrix, matching non-identical residue types can give positive scores as well. A bias of one residue type (e.g. arginine) can lead to high scores for a biased protein of a similar but not identical residue type (e.g. lysine). To prevent that unnecessarily many residues are blanked out due to such cross dependencies, we perform the masking iteratively. In each iteration, only the residue type with the highest score is masked, in the corresponding part of

```
Query: 1   MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV 60
           MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV
Sbjct: 1   MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV 60

Query: 61  LCAAGLFVFPIRGLISCLFHKSFQGCSGYVXXXXXXXXXXXXXIVGIVSCITWAPGFIFP 120
           LCAAGLFVFPIRGLISCLFHKSFQGCSGYV             IVGIVSCITWAPGFIFP
Sbjct: 61  LCAAGLFVFPIRGLISCLFHKSFQGCSGYVLATFLSLFSLALTIVGIVSCITWAPGFIFP 120

Query: 121 MISVSIAFATVETCFQIYTHLFPALEHKPXXXLKIEIAAAKLPRXXXAPDLNYPXLPTQX180
           MISVSIAFATVETCFQIYTHLFPALEHKP   LKIEIAAAKLPR   APDLNYP LPTQ
Sbjct: 121 MISVSIAFATVETCFQIYTHLFPALEHKPSSSLKIEIAAAKLPRSSSAPDLNYPSLPTQS 180

Query: 181 AXPXQRFXA 189
           A P QRF A
Sbjct: 181 ASPSQRFSA 189
```

Fig. 2. Comparison of SEG and CAST masking. SEG masking for a hypothetical protein taken from the *Chlamydia trachomatis* genome (gi|3328394), represented in the Blast output format. The query and subject sequences are identical, so that the biased region can easily be marked out. The segment at positions 91–103 (green) has been detected as biased, and filtered by replacing all residues with X residues. CAST detecting and masking a biased region for the same protein. Bias detection is not performed in a limited width window, often revealing biases that are spread throughout whole sequences. The detected segment (red), with bias caused by the high presence of serine residues (S) is masked in a more discriminating way: only serine symbols are masked, while the rest of the sequence information remains intact.

the sequence, provided that the score exceeds the chosen threshold (cut-off). Cross-scoring of similar residues to the masked one is avoided by this modification. The dynamic programming sweep is run on the masked sequence to detect if there are still biased regions with significant score to be masked. Continuous cycles of masking and detection are run until no more biased region scores higher than the cut-off.

The CAST algorithm has been implemented as a program that identifies and masks composition biased regions of which the score exceeds a given threshold (Figure 3). By default, it uses a 40 half-bit threshold, a value optimized for BLAST homology searches. A variant of BLOSUM62 (Henikoff and Henikoff, 1992) serves as the default scoring matrix, calculating the scores for 'X' as the mean value of the similarity scores in each row or column, as previously proposed (Altschul *et al.*, 1994), to eliminate the effect of the 'neutral' masking character in the next bias detection iteration. Although this scoring matrix performs well for general comparison purposes, other comparison matrices and various threshold values for the scores may be optionally chosen, as parameters.

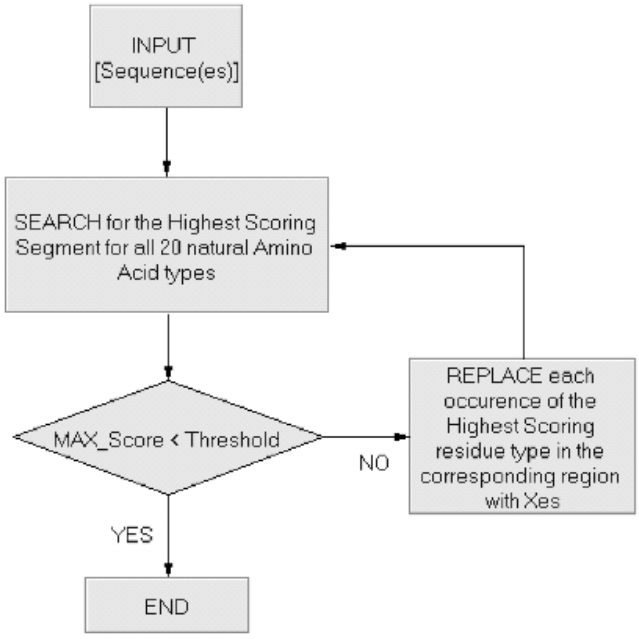

Fig. 3. Flowchart diagram of the CAST algorithm.

Results

Speed benchmarks

For a fixed set of program parameters (threshold value, mutation matrix), it was obviously expected that the time performance should mainly depend on the total number of the examined amino acids and the number of the sequences (as some overhead is added due to standard calculations for all sequences). It should be noted that the number of iterations cannot be predicted beforehand, and the performance of the algorithm cannot be evaluated in terms of a number of parameters, but only by simulation.

The composition of the tested sequences is critical, as the number and total size of reported biased segments is involved in the main output process, where besides the filtered sequence, a description of the bias type and the borders of the biased region are reported separately. While a change in the scoring matrix does not appear to drastically affect the speed of the program, a change in the significance threshold has a direct effect in the time needed to run CAST on a specific set of sequences. The lower the threshold, the more compositional bias is reported for the same sequence set.

For the needs of speed benchmarking, 1500 randomly selected sequences from Swiss-Prot served as the test set. They were randomly split in five individual subsets of 100–500 sequences. The CAST algorithm was executed on all these sets using the default comparison matrix (BLOSUM62) for threshold values 40 (default) and 30. The execution time is linear with respect to the length of the query sequence (not shown). Lowering the threshold, significantly increases computation time (a table with a speed benchmark is available on the web site as additional material).

Analysis of the P. falciparum *chromosome 2*

In order to estimate the performance and reliability of the CAST algorithm in the detection and masking of low-complexity segments, extensive tests have been run on a large number of protein sequences. A systematic analysis of all 210 translated open reading frames[†] of *P. falciparum* chromosome 2 (Gardner *et al.*, 1998) is reported. This set was chosen as a complete set of biased protein sequences from a eukaryotic chromosome. The choice was also influenced by our previous comparative study of sequence annotation (Tsoka *et al.*, 1999), utilizing similarity-based prediction methods. The results have been obtained employing CAST with the following default parameters:

- A standard BLOSUM62 comparison matrix, which is reported to perform best for database searching (Henikoff and Henikoff, 1993, 1996) and
- A significance threshold of 40 half-bits, a value optimized for BLAST homology searches (Altschul *et al.*, 1994).

For 156 out of the 210 sequences (approximately 74%) CAST identified at least one region of significant bias (Figure 4). A total of 547 biased regions were detected, with a mean value of approximately 3.5 regions per biased sequence (taking into account just the 156 sequences that contained at least one such segment), or 2.6 regions per sequence against the whole set. The maximum number of low-complexity segments in a single sequence was 18 (overlaps are allowed throughout this study).

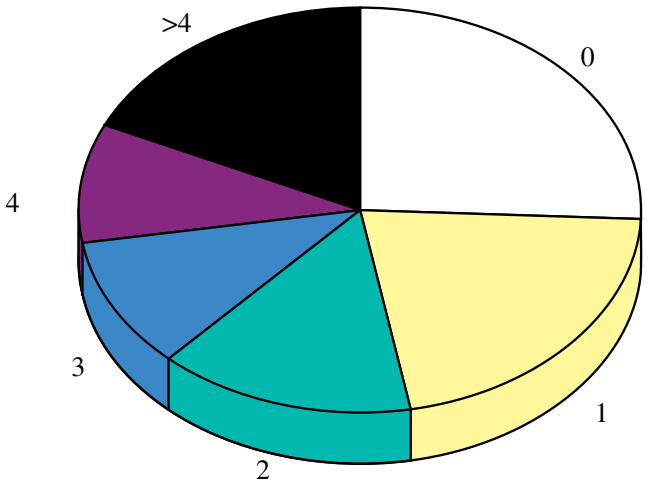

Fig. 4. Distribution of detected low-complexity regions per protein sequence for all 210 open reading frames of *P. falciparum* chromosome 2. One quarter of the ORFs have no such regions, while almost another quarter have four or more low-complexity runs. Amino acid residue type not shown.

Incorporating the information about the strand where each gene was coded in, we observed that 85 out of 106 genes (∼80%) present at the sense strand (defined as 'Crick') had at least one biased region, summing up to 295 detected segments (3.4 segments per biased sequence). The two sequences that had the largest number of biased segments (PFB0015 and PFB0020c, with 18 and 17 segments respectively) belong to this set. Examining the genes present at the complementary strand (defined as 'Watson') showed that 71 out of the remaining 104 sequences (∼68%) had at least one biased region, with a total of 252 biased regions (3.5 segments per biased sequence) (a table is available on the web site).

Per residue statistics are easy to obtain, as each biased region corresponds to a specific amino acid type (Figure 5). 300 out of the 547 identified regions (∼55%) were associated with Asparagine (N, 155 regions) and Lysine (K, 145 regions). This observation is somewhat expected, as the *P. falciparum* chromosome 2 is reported to have a base composition of 80.2% A + T (Gardner *et al.*, 1998). A *Plasmodium*-specific protein family (described as 'Repetitive Interspersed Family RIF-1' by the original authors) also appears to contribute to this effect (rich in Isoleucine: codons ATT, ATC, ATA). Therefore, as these proteins appear in distinct clusters on the chromosome, another interesting question is the occurrence of biased regions along the chromosome. All 20 possible bias types have been studied but no clear patterns emerge (not shown).

[†] The public ORF collection includes PFB0165w, which encodes tRNA-Glu.

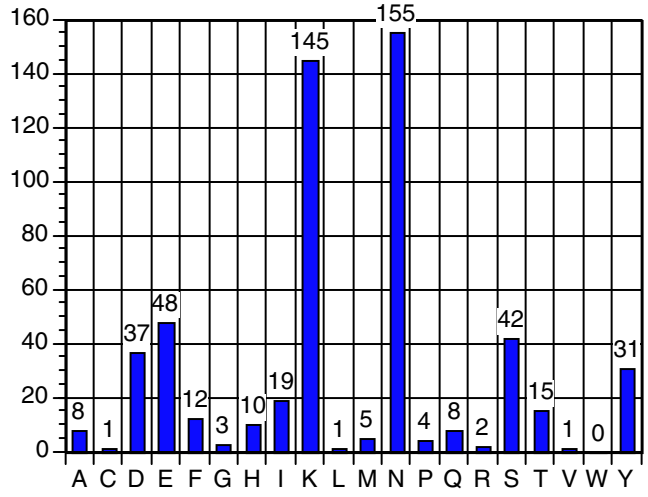

Fig. 5. Absolute counts of low-complexity regions of all 210 open reading frames of *P. falciparum* chromosome 2 protein sequences detected by CAST (*y*-axis), classified for the 20 amino acid residue types (*x*-axis). It is evident that lysine and asparagine-rich regions dominate.

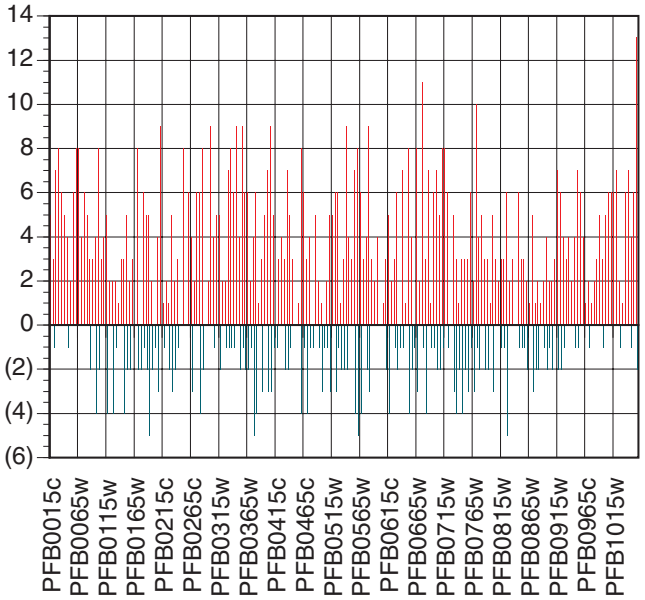

Fig. 6. Demonstration that detection of low-complexity regions by CAST stems from local complexity and not global composition bias. On the *x*-axis, the 210 open reading frames of *P. falciparum* chromosome 2 are listed (labels every ten ORFs). On the *y*-axis, the number of amino acid residue types detected by CAST only (red bars) or by both CAST and a global bias (counts in parentheses, dark green bars). Maximum number of residue types is 20. Significant global composition bias per residue was obtained for all ORFs, by comparing the observed residue composition of each ORF with the average residue composition of Swiss-Prot, Release 38: a chi-square test at 99.5% confidence level with 19 degrees of freedom (chi-square value = 38.6) was used.

Positional analysis of the occurrence of biased regions along the sequences of *P. falciparum* chromosome 2 has been performed for all residue types. As in some cases the sample is too small to obtain meaningful statistics, e.g. residue types involved in few or even no biased regions at all (like W as shown in Figure 5), results are only presented for the five most abundant bias types (N, K, E, S and D). These residues are responsible for 427 biased regions, almost 80% of the total. We normalised the start, middle and end positions of the biased segments as portions of the total length for each examined sequence (a table with the positional analysis of the five most abundant residue bias types is available on the web site as additional material). These results provide some useful knowledge to the study of biased regions as a stand-alone phenomenon, apart from the practical application of low-complexity detection for query masking in homology searches. The lengths of the biased regions seem to vary, as well as their locations along the sequences they appear in. For example the 37 D-rich biased regions in the test set tend to be in the C-terminal regions (not shown). Such observations could reveal possible structural and functional characteristics for proteins of unknown function. We have permitted overlaps of biased regions in this work so that each type of bias could be studied as a separate fact. A study of the possible combinations of different bias types in a sequence is currently in progress (in preparation).

Finally, we have addressed the issue of whether the detection of bias regions arises from the global amino acid composition of the test sequences or the effect of local low complexity. It appears that despite the unusual global composition of some sequences for certain residue types, the principal effect of detection originates from local bias composition (Figure 6).

Comparison to other methods

CAST has also been compared to the widely used SEG method. The set of 210 ORFs of *P. falciparum* chromosome 2 served as the test set to obtain some statistics on the performance of both methods in detecting and masking biased regions. In our comparison, the analysis with SEG was performed with the default parameters optimized for low-complexity masking of many amino acid sequences, as described by the authors in the distribution package (Wootton and Federhen, 1993): a trigger window length (W) of 12, a trigger complexity ($K_2(1)$) of 2.2 bits and an extension complexity ($K_2(2)$) of 2.5 bits. SEG reports composition bias in almost 90% of the sequences (188 out of 210), with a slight tendency to over-detect biased regions compared to CAST (156 out of 210 sequences, or 74%). A total number of 1321

biased regions are reported by SEG, with a mean value of approximately seven biased regions per biased sequence.

Another comparison between SEG and CAST involved the filtering of all the 210 ORFs of *P. falciparum* chromosome 2 used as queries in database searches (Tsoka *et al.*, 1999). For each ORF, we created four differently masked variants by applying either SEG or CAST with low and high stringency parameters. Each of these variants was then used to search the non-redundant protein sequence database using BLASTP (Altschul *et al.*, 1997). We then compiled an overview of common hits, comparing rank and significance of each hit for the four differently masked query variants. We also listed any additional hits only returned by individual variants. We further recorded some statistics and the functional assignments as previously reported (Tsoka *et al.*, 1999). The results of this analysis using E-value cut-off scores of 10^{-30} and 10^{-06} are available at the CAST web site.

It is noteworthy that there is a large number of ORFs where all four masked variants returned the same set of homologues (occasionally with slightly different E-value ordering). This is particularly striking for the stringent cut-off E-value of 10^{-30}, where 143 out of 210 ORFs (68%) are insensitive to the masking strategy employed. Even for the more relaxed cut-off E-value of 10^{-06}, for 63 out of 210 ORFs (30%) the choice of masking algorithm and parameters made no difference in the sets of homologues retrieved. We manually examined the differences for the other 147 ORFs of the latter search: roughly half of them show marginal differences only.

However, where large discrepancies were observed, queries masked by SEG generally returned many more hits below the E-value cut-off score. Often, the difference in numbers was striking: for example, PFB0765w, when masked by SEG yielded 139 hits while when masked by CAST returned only itself, agreeing with its assignment as 'unique' sequence (Tsoka *et al.*, 1999). In another example, PFB0335c when masked by SEG returned 83 hits versus 48 hits when masked by CAST. The extra hits found by SEG contained a much higher ratio of apparently unspecific matches, due to excessive filtering. There were only six sequences which, when masked by CAST, returned more hits than when masked by SEG (PFB0130w, PFB0215c, PFB0290c, PFB295w, PFB0410c and PFB0695c). Many of the additional hits for this set appear to be genuine homologues (result summaries may be downloaded from the CAST web site). In summary, CAST appears to allow more specific database searches without sacrificing sensitivity.

Searching databases for low-complexity regions

CAST has been developed to detect and mask unusual sequence composition in a single protein sequence by comparing it against a database of homopolymers composed of a distinct amino acid type each. Following this concept, a single homopolymer from this database can be compared to a complete database of natural proteins. Such a database search readily identifies the set of natural proteins that have composition bias of the corresponding residue type. This search can easily be performed for all 20 homopolymers and adequately reveals all kinds of biased regions existing in any sequence database, while at the same time biased regions can be classified by residue type. This database scan can be performed by any available homology search program. Similarly to the example of the 'poly-R' sequence against nrdb shown in Figure 1, such searches reveal intriguing patterns in the protein database. These patterns usually appear in a regular repetitive manner and may reflect some structural and functional principles.

In the command line implementation of CAST, an automatic search of a flat database file in FASTA format against the bias database can be performed. A next step would be to perform homology searches of filtered query sequences against a CAST-filtered database and optimize the CAST parameters for a better performance of well-tested, currently existing homology search methods.

Discussion

The CAST method is a novel low-complexity detection method that can be used both for masking query sequences for further analysis and the study of single amino acid residue types in protein sequences.

CAST is an indirect descendant of the '`biasdb`' program that has been extensively used in the GeneQuiz project (Andrade *et al.*, 1999), and allowed database searches with much lower significance cutoff values. One major improvement over `biasdb` is the treatment of the calculation of the 'X' residues as the mean value of the similarity scores.

A similar algorithm based on self-comparison has been reported (Marcotte *et al.*, 1998), that primarily aims to detect repeats in protein sequences. Some of the short-range repeats may indeed represent low-complexity regions but that problem was not further addressed.

CAST represents a useful method that can be used either for case-based sequence analyses to assist experimental biologists with sequence query filtering or massive sequence comparison for bioinformatics research.

Acknowledgements

This work was supported by the European Molecular Biology Laboratory and the TMR Programme of the European Commission (Directorate General-XII Science, Research and Development). C.O. thanks IBM Research for additional support. Earlier contributions by Georg Casari and other members of the GeneQuiz core team are gratefully acknowledged.

References

Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.

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.

Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.

Claverie,J.-M. and States,D.J. (1993) Information enchancement methods for large scale sequence analysis. *Comput. Chem.*, **17**, 191–201.

Gardner,M.J., *et al.* (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, **282**, 1126–1132.

Henikoff,J.G. and Henikoff,S. (1996) Blocks database and its applications. *Meth. Enzymol.*, **266**, 88–105.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.

Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1998) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.*, **183**, 63–98.

Robison,K., Gilbert,W. and Church,G.M. (1994) Large-scale bacterial gene discovery by similarity search. *Nature Genet.*, **7**, 205–214.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tsoka,S., Promponas,V.J. and Ouzounis,C.A. (1999) Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case. *FEBS Lett.*, **451**, 354–355.

Wootton,J.C. (1994) Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.*, **4**, 413–421.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.