[33] Analysis of Compositionally Biased Regions in Sequence Databases

By JOHN C. WOOTTON and SCOTT FEDERHEN

Introduction

Sequences of natural macromolecules are very different from random polymers, most strikingly in the numerous interspersed "simple" sequence regions that have significant biases in amino acid or nucleotide composition. From systematic analyses that use a conservative definition of significant bias, such regions account for approximately one-quarter of the amino acids in current sequence databases.[1,2] These include segments usually described by terms like glutamine-rich or glycine–arginine-rich, and also more weakly repetitive nonglobular domains. More than one-half of the proteins have at least one such region. Genomic DNA sequences are considerably more variegated than amino acid sequences, exhibiting many types of bias in different functional subclasses of sequences.[3]

Collectively, these regions exhibit a very broad range of compositional properties and lengths, and most of them have unknown structures, dynamics, and interactions.[4] Unprecedented classes keep on appearing in new genomic sequences and their coding sequence translations. The sequence simplicity varies from extreme, as in homopolymeric tracts, to very subtle as in some nonglobular domains of proteins. Locally abundant residues may be contiguous or loosely clustered, irregularly spaced or periodic. They tend to evolve rapidly, reflecting mutational processes such as replication slippage, unequal crossing-over, and biased nucleotide substitution. The relative roles of functional selection and mutational drive are generally unknown.

Therefore, lacking a consistent conceptual framework based on structure and evolution, computer analysis of low-complexity regions presents very different challenges than pairwise or multiple sequence alignment. In most cases it does not make sense from either structural or mutational

[1] J. C. Wootton and S. Federhen, *Comput. Chem.* **17,** 149 (1993).
[2] J. C. Wootton, *Comput. Chem.* **18,** 269 (1994).
[3] P. Salamon and A. K. Konopka, *Comput. Chem.* **16,** 117 (1992).
[4] J. C. Wootton, *Curr. Opin. Struct. Biol.* **4,** 413 (1994).

viewpoints to attempt to align low-complexity sequences position by position. Instead, general methods are required to explore and analyze nonrandom compositional heterogeneity in sequence databases. We have developed algorithms for these purposes, which are implemented in the well-tested SEG family of programs.[1,2] We provide here a practical guide to using these programs, and also a brief outline of their basic, but relatively unfamiliar, underlying principles.

Programs SEG and PSEG are tuned for amino acid sequences and NSEG for nucleotide sequences. The programs can be applied to either (1) individual sequences, including whole chromosomes if appropriate, or (2) entire sequence databases. The low-complexity regions can be defined for further study in their own right,[2,4] or, alternatively, can be masked from query sequences for sequence similarity searches in order to focus attention on alignments of high-complexity regions.[5]

The SEG philosophy is to use unbiased inference in an exploratory spirit. A sequence or database is treated initially as a heterogeneous mixture with unknown statistical properties, and then attempts may be made to infer these properties. An initial assumption of equal uniform probabilities for the appearance of residues places all possible low-complexity segments on an equal footing. For example, polypeptide segments rich in generally common amino acids such as leucine and alanine are treated as no more or no less surprising than segments rich in histidine, methionine, or tryptophan. This approach is justified by the results: in current databases, clusters of leucine are relatively rare, whereas segments rich in, for example, histidine or methionine are relatively abundant.

Having identified low-complexity regions at a given level of significance, we would then like to answer questions like, What does the segment resemble, and do any similar segments have known functions? This research requires database-oriented methods for segment comparison and classification, based on compositional attributes. These methods are under development[6] and are omitted from this chapter, as is a large body of relevant and fascinating mathematical theory, much of which has been reviewed by Konopka[7] under the name Biomolecular Cryptology. At the time of writing, only a few of these possible theoretical approaches have been implemented in robust software that is generally available via Internet. In addition to

[5] S. F. Altschul, M. Boguski, W. Gish, and J. C. Wootton, *Nat. Genet.* **6,** 119 (1994).

[6] J. C. Wootton and S. Federhen, *in* "Bioinformatics and Genome Research" (H. A. Lim and C. R. Cantor, eds.), p. 159. World Scientific Publishing, Singapore, 1995.

[7] A. K. Konopka, *in* "BIOCOMPUTING: Informatics and Genome Projects" (D. Smith, ed.), p. 119–174. Academic Press, San Diego, 1994.

the SEG family, these include SAPS,[8] XNU,[9] PYTHIA,[10] and SIMPLE34.[11] A comparative review of these and other methods has been published separately.[12]

Definitions

The SEG programs use general measures[1,3,13] of the combinatorial complexity of sequences and their compositions. Terms like "low-complexity" and "local compositional complexity" are used here only in this specific sense, distinct from the complexity theory of physicists, from algorithmic complexity, and from sequence complexity of classical reassociation kinetic experiments. We try to avoid the terms "entropy" and "information content" which have suffered from inconsistent usage in the biological literature.

Compositional complexity is based only on residue composition, regardless of the patterns or periodicity of sequence repetitiveness. This contrasts with some alternative methods[11,14,15] that use counts of k-grams (k-letter words; e.g., TAGG is a 4-gram) to define residue patterns and clustering. Complexity, pattern, and periodicity are distinct abstract attributes of simple sequences. For example, the following sequences have identical (low) compositional complexity because of their identical compositions (A_8T_8), but differ in patterns and periodicity:

TATATAAATTTAATTA has neither significant pattern nor periodicity
TAATTAATTTTAATAA has notable k-gram patterns (TAA, AAT, and TAAT) but these are irregularly spaced and do not show periodicity
TATATATATATATATA has periodicity, modulo-2, and hence significant k-gram patterns as a consequence

In the SEG programs, the complexity state of a sequence or subsequence is represented by a list of numbers, a complexity state vector,[1,3,13,16] which summarizes the composition. For example, a 5-nucleotide window has six

[8] V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, and S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 2002 (1992).
[9] J.-M. Claverie and D. J. States, *Comput. Chem.* **17,** 191 (1993).
[10] A. Milosavljevic and J. Jurka, *Comput. Appl. Biosci.* **9,** 407 (1993).
[11] J. M. Hancock and J. S. Armstrong, *Comput. Appl. Biosci.* **10,** 67 (1994).
[12] J. C. Wootton, *in* "Nucleic Acid and Protein Sequence Analysis: A Practical Approach" (M. J. Bishop and C. J. Rawlings, eds.), in press. IRL Press, Oxford, 1995.
[13] P. Salamon, J. C. Wootton, A. K. Konopka, and L. K. Hansen, *Comput. Chem.* **17,** 135 (1993).
[14] P. A. Pevsner, M. Y. Borodovsky, and A. A. Mironov, *J. Biomol. Struct. Dyn.* **6,** 1013 (1989).
[15] S. Pietrokovski, J. Hirshon, and E. N. Trifonov, *J. Biomol. Struct. Dyn.* **7,** 1251 (1990).
[16] A. K. Konopka and J. Owens, *Genet. Anal. Tech. Appl.* **7,** 35 (1990).

possible complexity state vectors. In order of increasing compositional complexity, these are {5,0,0,0}, {4,1,0,0}, {3,2,0,0}, {3,1,1,0}, {2,2,1,0}, and {2,1,1,1}. These vectors are ranked lists of the numbers of each nucleotide, irrespective of which letter corresponds to each number. The complexity state {3,1,1,0}, for example, has 12 different possible compositions, which include (T_3,C,A) and (G_3,T,C). Each of these compositions has 20 possible sequences, or permutations, of the letters in the composition.

Each complexity state has the same number of sequences per composition. The possibility of 20 sequences per composition makes {3,1,1,0} a more complex state than, for example, {3,2,0,0} which has only 10 possible sequences per composition. The theoretical number of complexity state vectors, which is computable from well-established principles of number theory, becomes very large at longer windows.[1,2] For example, an amino acid window of length 40 generates 35,251 complexity states and a rounded total of 1.1×10^{52} sequences.

Methods

We first describe the SEG program and its various applications in analyzing interspersed low-complexity amino acid sequences. Then we consider additional methods implemented in PSEG and NSEG to analyze periodic compositional complexity and nucleotide sequences.

Segment Representation

Low-complexity segments in natural protein sequences are most commonly interspersed, so that compositional bias may be considered to be a local property of contiguous residues.[1] These local contrasts in complexity may be represented by optimized segments, that is, nonoverlapping subsequences with precise boundaries that are defined as either high complexity or low complexity. The number of such segments and their lengths are determined automatically by the optimization algorithm in SEG, subject to parameters (which may be user-specified) that control the granularity and stringency of segmentation.

The optimized segments produced by SEG are represented, in either human-readable or machine-readable forms, as digitized text strings, together with their sequence position coordinates and other data. This approach is distinct from graphical methods such as dot-matrix plots or complexity profiles, which also have their uses in representing compositional bias (see below) but are not readily applied to large sequence databases. SEG can analyze sequences of any length, and also entire sequence databases. The segmented sequence strings, or their sequence identifiers and

coordinates, may be used as input for further computer analysis or stored as a specialized minidatabase.

Using SEG Program

The SEG program is run from a UNIX command line that specifies the input file. This file is an amino acid sequence, formatted as for the FASTA[17] or BLAST[18] programs, or a database file containing many such sequences. Using the example of the human prion protein (file "prion," Fig. 1A), the minimal command line for SEG is "seg prion" or, to direct the standard output to a file, "seg prion > prion.segout." This command implies the default parameters and options, as discussed below. The program runs in less than a second on typical UNIX workstations with sequences of up to approximately 1000 amino acids.

The default output is tree form (Fig. 1B), which is designed to be human-readable and to focus attention on segments of interest. Regions of contrasting complexity are displayed on the two sides of a central column of residue numbers, with low-complexity segments in lowercase letters on the left-hand side and high-complexity segments in uppercase letters on the right. The sequences of all segments read from left to right, and their position in the sequence is ordered top to bottom, as the residue numbers indicate. The sequence identifiers and definition line are also presented.

Some options of SEG give forms of output that are designed primarily for input to other computer programs. Commonly used examples are shown in Fig. 1C,D; a complete list of options is given in the documentation file available with the source code (see section on Availability below). The "-x" option (Fig. 1C) produces a masked FASTA-formatted file, ready for input as a query sequence for database search programs such as BLAST or FASTA. The amino acids in low-complexity regions are replaced with "x" characters. This option is implemented as a filter in the BLAST series of programs when the "-filter seg" option is specified.[19] Masking removes the potential confusion caused by low-complexity segments in database search methods such as BLAST or FASTA. Compositional bias is not encompassed by the random model used in these methods to evaluate local alignment statistics, and it can result in spuriously high scores and overwhelmingly large output lists.[5] One essential application of masking

[17] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444 (1988).

[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215,** 403 (1990).

[19] T. L. Madden, R. L. Tatusov, and J. Zhang, this volume [9].

(A) Input File (FASTA Format)

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C).
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
ILLISFLIFLIVG
```

(B) seg prion

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C).

                                   1-49   MANLGCWMLVLFVATWSDLGLCKKRPKPGG
                                          WNTGGSRYPGQGSPGGNRY
ppqggggwgqphgggwgqphgggwgqphgg    50-94
                gwgqphgggwgqggg
                                  95-112  THSQWNKPSKPKTNMKHM
       agaaaagavvgglggymlgsams    113-135
                                  136-187 RPIIHFGSDYEDRYYRENMHRYPNQVYYRP
                                          MDEYSNQNNFVHDCVNITIKQH
                tvtttkgenftet     188-201
                                  202-236 DVKMMERVVEQMCITQYERESQAYYQRGSS
                                          MVLFS
             sppvillisflifliv     237-252
                                  253-253 G
```

(C) seg prion -x

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C).
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxTHSQWNKPSKPKTNMKHMxxxxxxxxx
xxxxxxxxxxxxxxxxRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHxxxxxxxxxxxxxxxDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSxxxx
xxxxxxxxxxxxG
```

(D) seg prion -l

```
>PRIO_HUMAN(50-94) complexity=1.92 (12/2.20/2.50)
ppqggggwgqphgggwgqphgggwgqphgggwgqphgggwgqggg

>PRIO_HUMAN(113-135) complexity=2.47 (12/2.20/2.50)
agaaaagavvgglggymlgsams

>PRIO_HUMAN(188-201) complexity=2.26 (12/2.20/2.50)
tvtttkgenftet

>PRIO_HUMAN(237-252) complexity=2.50 (12/2.20/2.50)
sppvillisflifliv
```

FIG. 1. SEG input and outputs. The sequence is the unprocessed human prion protein, SWISS-PROT accession number P04156, name PRIO_HUMAN [H. A. Kretzschmar, L. E. Stowring, D. Westaway, W. H. Stubblebine, S. B. Prusiner, and S. J. Dearmond, *DNA* **5,** 315 (1986)]. (A) FASTA-formatted input file. (B, C, and D) Different forms of output given by the command lines shown and described in the text. The structural and functional significance of the low-complexity segments of the prion protein is unknown.

has been for the database–database comparisons used to precompute neighbors for the *Entrez* retrieval system.[20]

Some SEG output options write the optimized segments as a multisequence, FASTA-formatted "minidatabase" file, suitable for a wide range of further computer analyses. These may contain only the low-complexity segments ("-l" option, Fig. 1D) or only high-complexity segments ("-h" option), or both ("-a" option). Each segment is a separate sequence entry with an informative header line. Complexity data (K_2 in units of bits, see below) and other program parameters are included in the header line. The "-l" option is particularly useful for research on the low-complexity segments of whole sequence databases such as SWISS-PROT.[21] The SEG outputs from the "-l" option at different stringencies then become input for further searches, classification, and statistical analyses.[2,4,6]

Parameters

The number of low-complexity segments and their lengths are determined automatically by the optimization algorithm in SEG. However, the granularity and stringency of the search for low-complexity segments are controlled by three numeric parameters, which may be specified by the user on the command line. These are, in obligatory order after the sequence file name and before the options, W (trigger window length, an integer greater than zero, default 12 residues), $K_2(1)$ (trigger complexity, default 2.2 bits), and $K_2(2)$ (extension complexity, default 2.5 bits). The latter complexity values, for 20-letter amino acid sequences, may be greater than or equal to zero up to a maximum of 4.322 bits (the rounded value of $\log_2 20$).

The roles of these parameters are described in detail below in the section on the SEG algorithm. It is not necessary to understand the meaning of these parameters to use SEG creatively for many purposes, although knowledge of the algorithm is useful for more flexible parameterization and to help interpret unexpected results from database analyses. The default parameters are a good compromise for most proteins for the purpose of masking low-complexity regions, with the "-x" option, in query sequences for the BLAST programs. Thus, "seg prion -x" has the same meaning as "seg prion 12 2.2 2.5 -x." Other recommended standard parameter sets are as follows (for sequence file "myseq"): (1) for homopolymer analysis, to examine all homopolymeric subsequences of length 7 or greater, for example, use "seg myseq 7 0 0" (the complexity values of zero force maximum possible stringency, which identifies only homopolymers), and (2) for long

[20] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, this volume [10].
[21] A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **22,** 3578 (1994).

nonglobular domains of protein sequences, diagnose at longer window lengths,[2] for example, as discussed below, it is valuable to use two different granularities: "seg myseq 25 3.0 3.3" and "seg myseq 45 3.4 3.75."

Granularity

Figure 2 illustrates the concept of granularity, using both graphical complexity profiles and optimal segments produced by the SEG algorithm from the human prion protein sequence. In this example, the *W* (trigger window length) parameter was varied (5, 10, 20, or 30 residues). Longer trigger windows define more sustained regions of low complexity (compare 30-residue windows with those of 10 or 5, Fig. 2), but they overlook some

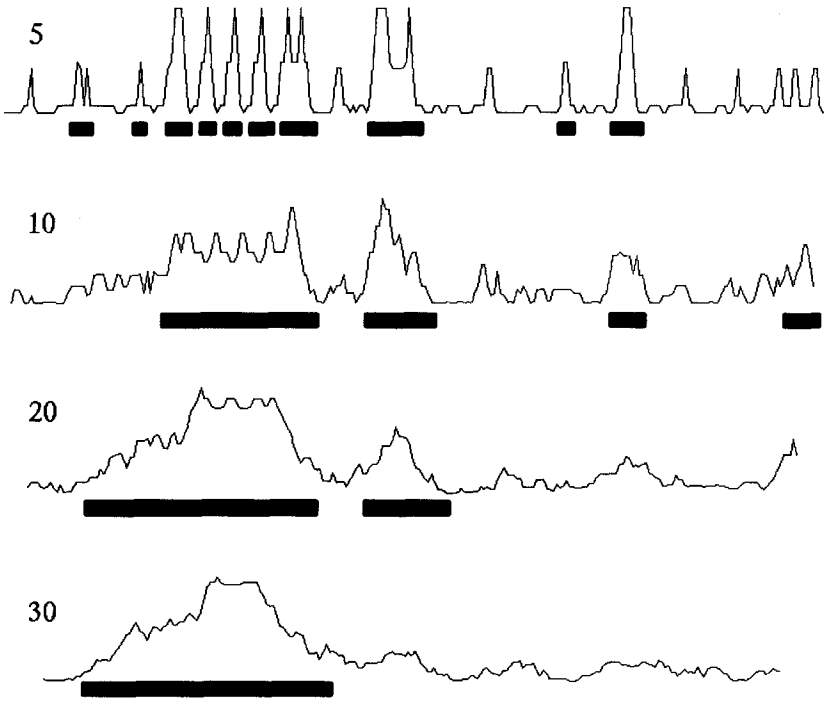

FIG. 2. Complexity profiles and optimized segments obtained from the human prion protein sequence (see Fig. 1 legend) at four different window lengths. The horizontal axis of the complexity profiles represents the 253 positions of the amino acid sequence. The vertical axis of each plot is the complexity, with low complexity at the top and high complexity at the bottom of each plot. The complexity measure is K_1 defined in the text section on the SEG algorithm, which is in the range 1–0. The solid bars below each plot show the extent of each optimal segment obtained in the corresponding four runs of SEG with the *W* parameter set to the same values as the window lengths of the profiles.

very short biased subsequences detected by the shorter windows. Thus, varying W changes the granularity, or resolution, of the search for low-complexity segments, but not necessarily the lengths of the optimized segments produced by SEG, which may be identical at different trigger window lengths.

Exploratory Strategy

The following is recommended as a routine strategy, to investigate a new protein sequence for possible low-complexity segments and nonglobular domains prior to database searches: (1) visual inspection of a self–self dot-matrix plot (many implementations are available), using a very low threshold, is recommended, so that low-complexity and repetitive regions appear as diagonal blotches and stripes (Fig. 3), providing a graphic, intu-

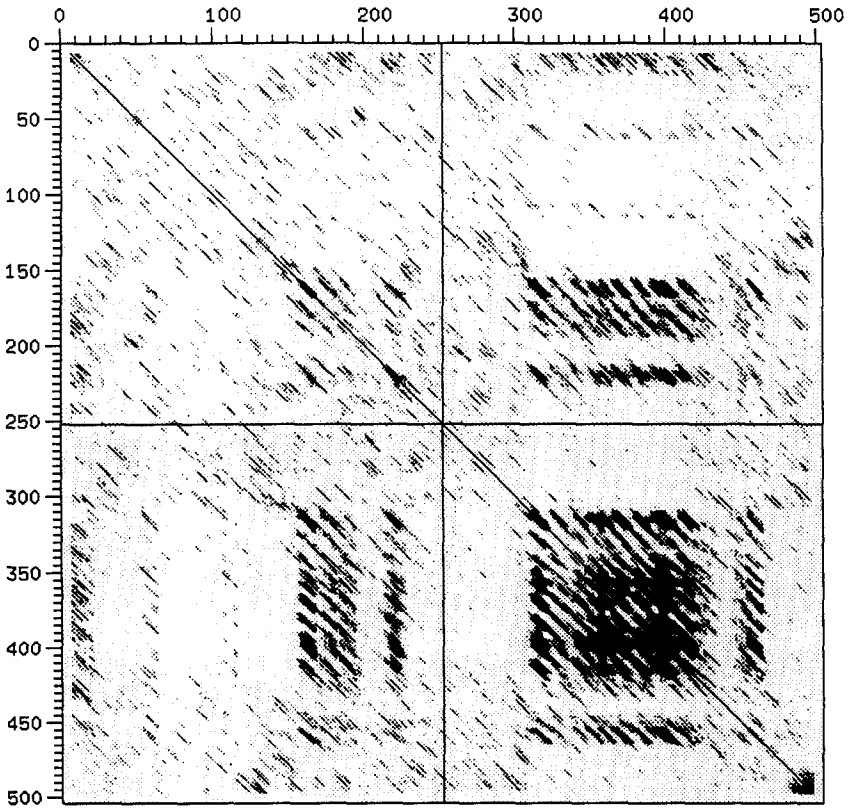

FIG. 3. A low-threshold, self–self dot-matrix plot of the human Wiskott–Aldrich syndrome protein (translated from GenBank accession number U12707, NCBI gi number 695151).

itive preliminary to more rigorous analysis using SEG; (2) SEG analysis with the default parameters and also the two sets (45 3.4 3.75 and 25 3.0 3.3) that target different length ranges of "nonglobular" regions is advised, as the default tree form output may focus attention on interesting regions and act as a guide to the use of masking in subsequent BLAST searches.

This strategy is illustrated in Figs. 3 and 4 for the human WASP gene product, which is mutationally modified in Wiskott–Aldrich syndrome, a hereditary immunodeficiency.[22] The protein is strikingly proline-rich, but the SEG results at different granularities show that a weakly repetitive pattern and possible nonglobular domains extend well beyond the obvious proline-rich segments. After masking of these potential nonglobular regions, a BLASTP output list then readily reveals possible sequence motifs in the high-complexity "globular" regions that weakly resemble parts of nucleolar transcription factor UBF-1. These features, and the potential nonglobular domains themselves, are candidates for further investigation by computer and experimental methods.

Nonglobular Domains

The advantage of using SEG to predict candidates for nonglobular domains of proteins is its generality.[2] Other, pattern-based methods may be used to classify specific classes of nonglobular structure, for example, types of coiled coil,[23] but the SEG method is sensitive to any regular or irregular weak repetitiveness of sequence, regardless of the actual polypeptide structure likely to be present.

Sequences of physicochemically defined nonglobular regions are partitioned accurately using the "nonglobular" SEG parameters recommended above, for example[2] (at $W = 45$), those of collagens, myosins, other coiled-coil proteins, elastins, mucins, proteoglycan core proteins, and long solvent-exposed α helices such as caldesmon, and (at $W = 25$) histones H1/H5 and nonhistone proteins. Candidates for new classes of unknown structure have also been predicted.[2] In the great majority of cases, these parameters appear to partition elongated domains, although intriguing exceptions are found in some regions of large heat-shock protein sequences that have subtle low-compositional complexity although they are known to possess a folded structure.

Periodic Compositional Complexity

Compositional bias commonly affects residues that are spaced at regular intervals rather than contiguous.[2,4,7] The concept of compositional complex-

[22] J. M. Derry, H. D. Ochs, and U. Francke, *Cell* (*Cambridge, Mass.*) **78,** 635 (1994).
[23] A. Lupas, M. Van Dyke and J. Stock, *Science* **252,** 1162 (1991).

(A) seg wiskott 25 3.0 3.3

```
>gp|U12707|HSU12707_1 Wiskott-Aldrich syndrome protein [Homo sapiens]

                                  1-147  MSGGPMGGRPGGRGAPAVQQNIPSTLLQDH
                                         ENQRLFEMLGRKCLTLATAVVQLYLALPPG
                                         AEHWTKEHCGAVCFVKDNPQKSYFIRLYGL
                                         QAGRLLWEQELYSQLVYSTPTPFFHTFAGD
                                         DCQAGLNFADEDEAQAFRALVQEKIQK
rnqrqsgdrrqlpppptpaneerrgglppl    148-200
       plhpggdqggppvgplslglatv
                                  201-204  DIQN
pditssryrglpapgpspadkkrsgkkkis   205-243
                     kadigapsg
                                  244-311  FKHVSHVGWDPQNGFDVNNLDPDLRSLFSR
                                         AGISEAQLTDAETSKLIYDFIEDQGGLEAV
                                         RQEMRRQE
plpppppppsrggnqlprapivggnkgrsgp   312-434
lppvplgiappppptprgppppgrgglhhhp
lqlldvldhcplhplelvghpchhhrhhrh
rrpapgmdqplphslllwclpgawpgggrg
                              all
                                  435-483  DQIRQGIQLNKTPGAPESSALQPPPQSSEG
                                         LVGALMHVMQKRSRAIHSS
            degedqagdededdewdd    484-501
```

(B) seg wiskott 45 3.4 3.75

```
>gp|U12707|HSU12707_1 Wiskott-Aldrich syndrome protein [Homo sapiens]

                                  1-116  MSGGPMGGRPGGRGAPAVQQNIPSTLLQDH
                                         ENQRLFEMLGRKCLTLATAVVQLYLALPPG
                                         AEHWTKEHCGAVCFVKDNPQKSYFIRLYGL
                                         QAGRLLWEQELYSQLVYSTPTPFFHT
fagddcqaglnfadedeaqafralvqekiq   117-259
krnqrqsgdrrqlpppptpaneerrgglpp
lplhpggdqggppvgplslglatvdiqnpd
itssryrglpapgpspadkkrsgkkkiska
       digapsgfkhvshvgwdpqngfd
                                  260-296  VNNLDPDLRSLFSRAGISEAQLTDAETSKL
                                         IYDFIED
qggleavrqemrrqeplpppppppsrggnql   297-469
prapivggnkgrsgplppvplgiappppptp
rgpppppgrgglhhhplqlldvldhcplhpl
elvghpchhhrhhrhrrpapgmdqplphsl
llwclpgawpgggrgalldqirqgiqlnkt
       pgapessalqpppqsseglvgal
                                  470-501  MHVMQKRSRAIHSSDEGEDQAGDEDEDDEW
                                         DD
```

FIG. 4. SEG results using searches of the human Wiskott–Aldrich syndrome protein (see Fig. 3) with two different granularities of "nonglobular" parameters. The SEG command lines are shown, with the parameters following the sequence file name "wiskott." (A) Low-complexity segments correspond approximately to the repetitive regions that are clear in Fig. 3. (B) Parameters with $W = 45$ reveal additional weakly biased sequences flanking the more striking proline-rich repeats, and these additional regions are recruited into consolidated low-complexity segments. For BLASTP searches, it is appropriate for this protein to mask these longer, consolidated segments, in order to focus attention on any weak similarities in the high-complexity, probably globular, regions.

ity is readily extended to these periodic cases, as implemented in programs PSEG for proteins and NSEG for nucleic acids.

Periodicity is a formal sequence attribute that is independent of compositional complexity and is defined as repetition of residue types or *k*-grams at a constant interval (period, modulo, or distance). It is useful to distinguish true periodicity, which is tandem repetition, exact or with variations, of a sequence pattern of constant length, from quasi-periodicity[7] in which repetition arises as a secondary consequence of different compositional biases in different phases. Quasi-periodic regions are abundant in DNA sequences, for example, modulo-3 in codon-biased mRNA coding sequences, and also occur in some helical polypeptides, for example, modulo-3 in collagen rods. Nonintegral periods may also arise from helical secondary structures in polynucleotide or polypeptide chains.

To generalize the concept of complexity state vectors to noncontiguous positions in a sequence, periodic compositional complexity is calculated from complexity state vectors whose numbers are counts of residues spaced at any defined interval. For example, a modulo-3 window of length 12 counts only the residues marked "1", and not the residues marked "0", in a 36-residue subsequence:

(1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0,1,0,0)

Sliding this window along a sequence in steps of single residues defines three overlapping phases. The period in PSEG is set by the "-z 3" option (for period 3). PSEG output uses lowercase characters, or "x" characters for masking with the "-x" option, to represent only those individual residues that are defined as low-complexity in a phase-specific manner (Fig. 5). With PSEG and NSEG, tree form output is specified by the "-p" command line option.

Figure 5 illustrates the phase specificity of PSEG results, using a typical collagen sequence. As with the "nonglobular" parameters of SEG, the sequence is partitioned accurately and automatically by PSEG into its globular domains (Fig. 5A, right-hand blocks) and nonglobular rods (Fig. 5A, left-hand blocks). However, PSEG has the important advantage that high-complexity information is retained within the rod sequences (shown by uppercase characters among the lowercase low-complexity phases). Masking with PSEG using the "-x" option (Fig. 5B) enables BLASTP to be used to investigate the subclasses and phylogeny of collagens from their rod sequences alone. The glycine phase of the rod segments and the proline-rich (including hydroxyproline) parts of the other two phases are defined by PSEG as low complexity, so that large output lists of spurious matches to glycine-rich or proline-rich noncollagen proteins are completely eliminated.

(A) pseg collagen -z 3 -p

```
>CA19_HUMAN COLLAGEN ALPHA 1(IX) CHAIN PRECURSOR.

                                    1-266  MKTCWKIPVFFFVCSFLEPWASAAVKRRPR
                                           FPVNSNSNGGNELCPKIRIGQDDLPGFDLI
                                           SQFQVDKAASRRAIQRVVGSATLQVAYKLG
                                           NNVDFRIPTRNLYPSGLPEEYSFLTTFRMT
                                           GSTLKKNWNIWQIQDSSGKEQVGIKINGQT
                                           QSVVFSYKGLDGSLQTAAFSNLSSLFDSQW
                                           HKIMIGVERSSATLFVDCNRIESLPIKPRG
                                           PIDIDGFAVLGKLADNPQVSVPFELQWMLI
                                           HCDPLRPRRETCHELPARITPSQTTD
ergppgeqgppgasgppgvpgidgidgdrg      267-356
pkgppgppgpagepgKpgApgKpgTpgAdg
LtgPdgSpgSigSkgQkgEpgVpgSrgFpg
rgIpgppgppgtaglpgelgrvgpvgdpgr      357-754
rgppgppgppgprgtIgFHdGDpLCpNAcP
PgRSgYPgLPgMRgHKgAKgEIgEPgRQgH
KgEEgDQgELgEVgAQgPPgAQgLRgITgL
VgDKgEKgARgLDgEPgPQgLPgAPgDQgQ
RgPPgEAgPKgDRgAEgARgIpgLpgPkgD
tgLpgVdgRdgIpgMpgTkgEpgKpgPpgD
agLqgLpgVpgIpgAkgVagEkgStgApgK
pgQmgNsgKpgQqgppgevgprgpqglpgs
rgelgpvgspglpgklgslgspglpglpgp
pglpgmkgdrgvvgepgpkgeqgaSgeEge
AgeRgeLgdIglPgpKgsAgnpgepglrgp
egsrglpgVegPrgPpgPrgVqgEqgAtgL
                          pgVqgPpg
                                    755-786  RAPTDQHIKQVCMRVIQEHFAEMAASLKRP
                                             DS
gATgLpgRpgPpgPpgPpgEngFpgQmgIr      787-899
gLpgIkgPpgAlgLrgPkgDlgEkgErgPp
gRgPNgLPgAIgLPgDPgPAsYgkNgrDge
       RgppglagIpgVpgPpgPpgLpg
                                    900-931  FCEPASCTMQLVSEHLTKGLTLERLTAAWL
                                             SA
```

(B) pseg collagen -z 3 -x

```
>CA19_HUMAN COLLAGEN ALPHA 1(IX) CHAIN PRECURSOR.
MKTCWKIPVFFFVCSFLEPWASAAVKRRPRFPVNSNSNGGNELCPKIRIGQDDLPGFDLI
SQFQVDKAASRRAIQRVVGSATLQVAYKLGNNVDFRIPTRNLYPSGLPEEYSFLTTFRMT
GSTLKKNWNIWQIQDSSGKEQVGIKINGQTQSVVFSYKGLDGSLQTAAFSNLSSLFDSQW
HKIMIGVERSSATLFVDCNRIESLPIKPRGPIDIDGFAVLGKLADNPQVSVPFELQWMLI
HCDPLRPRRETCHELPARITPSQTTDxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxKxxAxxKxxTxxAxxLxxPxxSxxSxxSxxQxxExxVxxSxxFxxxxIx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxIxFHxGDxLCxNAxPPxRS
xYPxLPxMRxHKxAKxEIxEPxRQxHKxEExDQxELxEVxAQxPPxAQxLRxITxLVxDK
xEKxARxLDxEPxPQxLPxAPxDQxQRxPPxEAxPKxDRxAExARxIxxLxxPxxDxxLx
xVxxRxxIxxMxxTxxExxKxxPxxDxxLxxLxxVxxIxxAxxVxxExxSxxAxxKxxQx
xNxxKxxQxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxSxxExxAxxRxxLxxIxxPxxKxxAxxxxxxxxxxxxxxx
xxxxVxxPxxPxxPxxVxxExxAxxLxxVxxPxxRAPTDQHIKQVCMRVIQEHFAEMAAS
LKRPDSxATxLxxRxxPxxPxxPxxExxFxxQxxIxxLxxIxxPxxAxxLxxPxxDxxEx
xExxPxxRxPNxLPxAIxLPxDPxPAxYxxNxxDxxRxxxxxxxIxxVxxPxxPxxLxxF
CEPASCTMQLVSEHLTKGLTLERLTAAWLSA
```

FIG. 5. Tree form (A) and masked (B) output from PSEG using the command line options shown. Other parameters are the defaults, namely, $W = 12$ residues, $K_2(1) = 2.2$ bits, $K_2(2) = 2.5$ bits. The sequence is human collagen alpha I type IX, SWISS-PROT accession number P20849, name CA19_HUMAN [Y. Muragaki, T. Kimura, Y. Ninomiya, and B. R. Olsen, *Eur. J. Biochem.* **192,** 703 (1990)].

"Simple" Sequences in DNA

Compositional bias in genomic DNA is much more intricate than in proteins, and it differs markedly in different organisms. DNA sequences are mosaics of different patches such as microsatellites, variable dinucleotide, trinucleotide, and other tandem repeats, dispersed repeats, telomeric sequences, recombinational hot spots, CpG islands, longer domains of different compositional bias, and many quasi-periodicities of different degrees of subtlety in both coding and noncoding sequences. In some cases, statistically distinct domains may overlap one another. This variegation cannot be adequately represented by simple random or Markov statistical models based on k-grams, and local low complexity is only one of the linguistic characteristics.[7,24]

The NSEG program is a useful discovery tool for exploring this difficult territory. The principles of NSEG are exactly as described above for SEG and PSEG, but tuned to the four-letter alphabet. The "-z" option specifies the periodicity. For example, "-z 1" defines a period of 1 and thus restricts the analysis to continuous segments of residues (as described above in the section on using the SEG program). Values of "-z" greater than 1 give behavior as described above for PSEG (see section on periodic compositional complexity).

The NSEG program is also effective for masking residues in biased regions prior to database searches by programs such as BLASTN that compare nucleotide sequences. The "-x" option in NSEG replaces masked nucleotides with "n" characters. The results of masking at several different periods can be integrated, using the utility program NMERGE to combine the "n" characters obtained from different NSEG output files and produce a single FASTA-formatted input file for BLASTN. A corresponding utility, PMERGE, combines different sets of "x" characters from amino acid sequences masked at different periods.

As with protein sequences, if the DNA sequence is not too long, a valuable preliminary step involves visual inspection of a self–self dot-matrix plot. For DNA, this should include matches of the sequence to its complementary strand. This plot may indicate the approximate positions and extent of repeats and compositional bias. For more rigorous identification of biased regions in DNA, which vary greatly in length, periodicity, and complexity, an interactive, exploratory approach using NSEG is recommended. This may be achieved by varying the command line parameters and investigating a range of periods, for example, from 1 to at least 6.

Possible parameters for an initial investigation might be, for period 1

[24] S. Karlin and V. Brendel, *Science* **259,** 677 (1993).

(contiguous residues), "nseg mychromosome 21 1.3 1.3 -z 1 -p" and, for periods of 2 or greater, more stringent parameters, for example, "nseg mychromosome 21 1.1 1.1 -z 3 -p." Shorter windows, for example 15 or 17 instead of 21, are likely to reveal additional short segments. Figure 6 shows a sample of NSEG results at periods 1 and 6 from part of a large intron of the human ABL gene, illustrating a purine-rich feature and a very subtle modulo-6 quasi-periodicity.

Following a few exploratory runs with different parameters, NSEG may then readily be used with the "-l" option to report, for example, all trinucleotide repeat segments in a whole genome or database at a required level of stringency.

SEG Algorithm

SEG is a two-pass algorithm. The first stage identifies approximate raw segments of low complexity. The second stage is local optimization within each raw segment. At the first stage, the search for low-complexity raw segments is determined by W, $K_2(1)$, and $K_2(2)$, described in the section on parameters above. $K_2(2)$ must not be less than $K_2(1)$ and is usually set slightly greater in order to permit the local optimization at the second stage of SEG to operate within a relatively liberal range.

The compositional complexity, K_1, of a complexity state vector (as used in the Fig. 2 profiles) is a measure of the information needed per position, given the composition of the window, to specify a particular residue sequence.[1,3] For an N-residue alphabet (usually, $N = 4$ or 20) and a window of length L:

$$K_1 = \frac{1}{L} \log_N \Omega \tag{1}$$

where Ω is the multinomial coefficient $(L!/\Pi_{i=1}^{N} n_i!)$, which gives the number of sequences per composition characteristic of the complexity state vector. The n_i are the N numbers in the complexity state vector. The logarithm is taken to base N to place K_1 in the range 0 to 1. To express complexity in frequently used information units, logarithms may be taken to base 2, giving bits, or to base e, giving nats.

Another informational measure of compositional complexity, K_2, expressed in bits, is used instead of K_1 for computational efficiency at the first stage of the SEG program:

$$K_2 = -\sum_{i=1}^{N} \frac{n_i}{L} \left(\log_2 \frac{n_i}{L} \right) \tag{2}$$

(A) nseg abl.1b 21 1.1 1.1 -z 1 -p

```
>U07562 Human ABL gene, intron 1b, subsequence 43040-43749.

                                    1-28   CCTGGGTGACAAAGTGAGACCCTATCTC
          aaaaaagaaagaaagaaaggaag  29-51
                                   52-60   TTAGAAAGT
          gggggagggagggaaggaggaag  61-83
                                   84-84   T
gaggaaggaaagaaaggaaagaaaacaaag  85-123
                        agaaagaaa
                                  124-710  CTTGTCTGATTTGGTCTGCACTTTGAGTGA
                                           GAGGAAACAAAAACCTCCCGAGATAACTTT
                                           TATTCTAAGCATGCCTTCTTTTAAACTTTG
                                           TTTGAAATTTACCTGGCCTCTGTAGTTTGTG
                                           AAATTCCAAGAAGAGACATTATTTTAGAGA
                                           GATACTAGGTTGCTAGAGTTCCTAGATTAC
                                           TACAGTTCCTAGTTAGCTGGCAGAAGTCAA
                                           TGCATATCCTCTTGGACTATATCTATCTTT
                                           TTCTCAGTCTTAACACCAGATTCCCAAAGA
                                           TAAAGCTTCTCCAAACATGATCTCACAATC
                                           CAAAACTAAGAACATGAGAAAACAAACAAC
                                           CTTGCATCACAGAGCAGGAAAATGAACAGT
                                           AGAACTATAGTACTTTATAGCTTTAGATAT
                                           AAAAACTGTCAGATTCAGAATATAAAATAC
                                           TGGTTTAAAAACATTTAAAGAAATAAAATG
                                           TGAAATAGAAAACAAGTAAGGACTAAAATC
                                           CTATGAAGAAAAATCAAATAATTGAAAAAG
                                           AACTAAAAAAGAAATGAAAACTACGATCAG
                                           TGAAATTAAAACTTCAGTGGATGGGTTTAA
                                           CAAGTTAGACACAACTG
```

(B) nseg abl.1b 21 1.1 1.1 -z 6 -p

```
>U07562 Human ABL gene, intron 1b, subsequence 43040-43749.

                                    1-441  CCTGGGTGACAAAGTGAGACCCTATCTCAA
                                           AAAAGAAAGAAAGAAAGGAAGTTAGAAAGT
                                           GGGGGAGGGAGGGAAGGAGGAAGTGAGGAA
                                           GGAAAGAAAGGAAAGAAAACAAAGAGAAAG
                                           AAACTTGTCTGATTTGGTCTGCACTTTGAG
                                           TGAGAGGAAACAAAAACCTCCCGAGATAAC
                                           TTTTATTCTAAGCATGCCTTCTTTTAAACT
                                           TTGTTTGAATTTACCTGGCCTCTGTAGTTT
                                           GTGAAATTCCAAGAAGAGACATTATTTTAG
                                           AGAGATACTAGGTTGCTAGAGTTCCTAGAT
                                           TACTACAGTTCCTAGTTAGCTGGCAGAAGT
                                           CAATGCATATCCTCTTGGACTATATCTATC
                                           TTTTTCTCAGTCTTAACACCAGATTCCCAA
                                           AGATAAAGCTTCTCCAAACATGATCTCACA
                                           ATCCAAAACTAAGAACATGAG
aAAACAaACAACcTTGCAtCACAGaGCAGG  442-652
aAAATGaACAGTaGAACTaTAGTAcTTTAT
aGCTTTaGATATaAAAACtGTCAGaTTCAG
aATATAaAATACtGGTTTaAAAACaTTTAA
aGAAATaAAATGtGAAATaGAAAAcAAGTA
aGGACTaAAATCcTATGAaGAAAAaTCAAA
tAATTGaAAAAGaACTAAaAAAGAaATGAA
                              a
                                  653-710  ACTACGATCAGTGAAATTAAAACTTCAGTG
                                           GATGGGTTTAACAAGTTAGACACAACTG
```

FIG. 6. Part of intron 1b of the human ABL gene, analyzed for contiguous-residue low-complexity segments (A) and quasi-periodicities of period 6 (B) using NSEG with the command lines shown. The subsequence indicated is from GenBank accession number U07562 (unpublished cosmid sequence submitted to GenBank by B. R. Roe, 1994).

K_2 is an approximation that converges toward K_1 at large window lengths.[1] K_2 is an adequate estimate of complexity for the first, noncritical, stage of SEG. K_2 is analogous to "entropy" in Shannon's information theory, whereas K_1, being based on combinatorial enumerations of states, resembles "entropy" in the sense of Boltzmann.

A sliding window of length W is moved in single-residue steps along the sequence, and complexity K_2 is computed at each step. Trigger windows are those of complexity less than or equal to $K_2(1)$ bits. A raw segment may consist of a single trigger window or a contig of overlapping windows. The contig is constructed by merging each trigger window in both directions with any overlapping trigger windows, and also with any overlapping extension windows of length W and complexity less than or equal to $K_2(2)$ bits. Each nonoverlapping contig is a raw segment.

At the second stage of SEG, each raw segment is reduced to a single optimal low-complexity segment, which may be the entire raw segment but is usually a shorter subsequence. The optimal subsequence is that with the most improbable composition, calculated on the assumption of equal (uniform) probabilities for the appearance of the 4 nucleotides or 20 amino acids. The probability of occurrence, P_0, of each complexity state is minimized:

$$P_0 = \frac{1}{N^L} \Omega F \tag{3}$$

where F is the combinatorial expression $N!/\Pi_{k=0}^{L} r_k!$, which is the number of compositions that have this complexity state vector.[1,3] Here, r_k is the count of the number of times the number k occurs among the n_i of the complexity state vector. Other symbols are as for formula Eq. (1). N^L is the total number of possible sequences for a window of length L and alphabet size N.

The segment of minimal P_0 is found by exhaustive search over all subsequences of each raw segment, using precomputed lookup tables of log factorials for efficiency. P_0 is particularly suitable for this optimization because it gives closely similar expected values at all window lengths.[1] The use of uniform prior probabilities of residues, as embodied in this probability function, is critical to ensure that low-complexity segments are defined in a consistent manner throughout a sequence or database.

Availability

The programs SEG, PSEG, NSEG, PMERGE, and NMERGE are available as standard C language source code that compiles on Sun, Solaris, or Silicon Graphics systems and has been compiled on other UNIX work-

stations with trivial modifications. Executables are also provided for Sun and Silicon Graphics systems. All these programs are available by anonymous ftp from ncbi.nlm.nih.gov in subdirectory /pub/seg together with documentation. A default version of SEG for masking purposes is also available as the "-filter seg" option within the BLAST programs.[19]

Future Developments

It is increasingly important to have robust methods for comparison and classification of compositionally biased segments of sequences. For a growing number, still a minority, of low-complexity protein segments, molecular interactions and functional assignments have been made experimentally. Given the involvement of such segments in biological functions such as morphogenesis and human molecular diseases, we would like to understand their crucial molecular characteristics that may be compositionally determined and generally evolve rapidly. Software to support such research is under development.[6]

For genomic studies, it is essential to view compositional bias in the context of many types of other features such as recognizable functional sites, transcripts, coding sequences, and homologies. For this purpose, the SEG family of programs is being integrated into software packages, or workbenches, that have graphic multilevel browsing facilities and include zoom functions. For example, graphical complexity profiles and SEG optimal segments can be viewed, at different granularities, in conjunction with sequence matches from BLAST programs in an enhancement of the BLIXEM viewer[25] that is under test at the time of writing.[26]

[25] E. L. Sonnhammer and R. Durbin, *Comput. Appl. Biosci.* **10,** 301 (1994).
[26] E. L. Sonnhammer and J. C. Wootton, unpublished (1995).