Sequences with 'unusual' amino acid compositions.

Review article

John C Wootton

Current Opinion in Structural Biology 1994, **4:** 413-421.

Outline

Abstract

Amino acid sequences of very non-random composition ('low-complexity' segments) are abundant in natural proteins. From recent statistical analyses of protein sequence databases, approximately 15% of the

residues occur in segments of extreme compositional bias, and approximately 34% of proteins have at least one such interspersed segment. Sequences of many elongated non-globular domains also have non-random compositional bias, and these regions increase the proportion of residues in statistically deviant segments to approximately 25% of the database. In contrast, less than 1% of residues in known ordered crystal structures are in segments of reduced complexity. Increasingly, low-complexity segments have been implicated in crucial biological functions, shown by genetic engineering and mutagenesis experiments, variations in human disease and locations of autoimmune epitopes, but relatively little is known about their range of possible molecular structures, dynamics and interactions.

Introduction

Segments of non-random amino acid composition, or 'low-complexity' regions, are very abundant in natural protein sequences. They include sequences rich in Ala, Gly, Pro, Gln, Ser, Thr, Asn, Glu, Asp, Arg, His, Met, Lys, Ile, Leu, Val, Phe residues, or combinations of a few of these. Some segments are homopolymers or nearly so, while others are irregular mosaics of mixtures of two or a more residues, and some include short-period regular repeats. They are strikingly abundant in large multidomain polypeptides crucial in morphogenesis, embryonic development, transcriptional regulation, RNA processing, signal transduction and both intracellular and extracellular structure and integrity.

Several hundred publications from 1993 and the previous few years (too many to list individually in this review) have reported new amino acid sequences, deduced from genomic or cDNA sequences, that contain low-complexity regions or domains. Almost all authors have used terms of surprise such as 'unusual', 'unexpected', 'extraordinary' and 'remarkable', perhaps reflecting a belief that such segments are rare, which they are not, or an expectation that normal proteins should have locally complex, quasi-random compositions.

The current protein sequence and structure databases [1][2][3][4] provide rich data for determining the actual abundance and nature of low-complexity segments using appropriate mathematical definitions of complexity. The statistical improbability of several specific classes of segments has been discussed for several years by Karlin and Brendel [5][6][7], with emphasis on clusters of charged amino acids. Recently, more general statistical measures of compositional complexity have been applied to both amino acid and nucleotide sequences and to entire sequence databases [8•][9][10][11], and these form the basis for the updated surveys reviewed here. Also reviewed is recent evidence implicating some of the low-complexity segments themselves in crucial molecular interactions and biological functions of these proteins. In only a few cases are relevant physicochemical details beginning to emerge, and there are many challenges for future research at the level of molecular structure and dynamics.

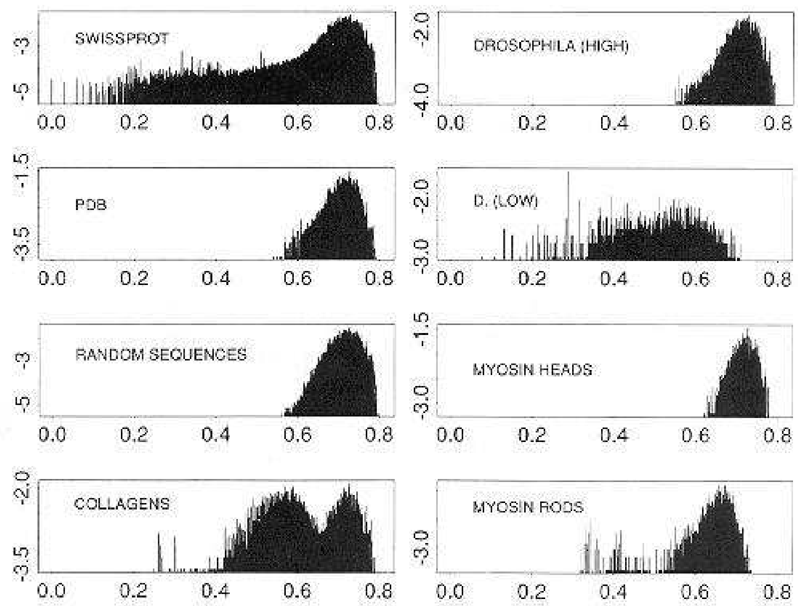

Fig. 1. Distribution of local compositional complexity in SWISS- PROT [2], PDB [4], random sequences of the same amino acid composition as SWISS-PROT, and various subsets of SWISS-PROT. The horizontal axis of each plot is Complexity K_1(see text) calculated for all windows of length *L*= 40 moved in steps of one residue along all the sequences in each data set. The vertical axis is $\log_{20}$(*frequency*), where *frequency*is the fraction of the total windows having each value of K_1. For *L*= 40, there are 35 251 possible complexity states. The *Drosophila*and myosin subsets of SWISS-PROT were partitioned into low-complexity and high-complexity segments using the SEG algorithm [8•] before calculation of the complexity distributions.

Return to Figure reference in text 1 , 2 , 3 , 4 , 5 , 6 , 7

Measurement of local compositional complexity

The method described below provides a measure of compositional complexity of a segment of sequence, defined as a general formal property independent of any periodic or irregular repetition of pattern [8•][9][10][11]. All possible segments are placed on an equal footing. For example, regions rich in generally common amino acids such as Leu, Ala and Ser are treated as no more or less surprising than segments rich in His, Met or Trp. In analyzing 'non-random' mosaics, we do not know what to expect, and in reality homopolymeric regions taken together have a very different amino acid composition from the whole sequence database [8•][11].

MASTERMIND – *Drosophila virilis*

LOW-COMPLEXITY SEGMENTS		HIGH-COMPLEXITY SEGMENTS
mdagglpvfqsasqaaaaavaqqqqqqqqqqqqqqqqqhlnlqlhqqhlqqqqslgih lqqqqqlqlqqqqqhnaqaqqqqqlqvqqqqqqrqqqqqqqqqhslynanlaaaggivgg lvpggngaggvalqqvfggpngnnnsnnnnnsnnnsininngnispgdglptkrqpildr lrrrmenyrrr	1-191	
	192-224	QTDCVPRYEQTFSTVCEQQNHETSALQKRFLES
knkraakktekklpetqqqaqtq	225-247	
	248-285	MLAGQLQSSVHVQQKILKRPADDVDNGAENYEPPQKLP
nnnnnnnnnnnnnnnnsssgvggs	286-309	
	310-355	ENLTKFSVEIVQQLEFTTSAANSQPQQISTNVTVKALTNTSVKSEP
gvgggrgrhqqqqqhqqhqqqqhqqqqhqqhqqhqqqqqhqqqqhqqqqhqqqqqqhhhq qqqqqgggglgglgnngrggggpggggghmatgpggvgvgmgpnmm	356-459	
	460-569	SAQQKSALGNLANLVECKREPDHDFPDLGSLAKDGANGQFPGFPDLLGDDNSENNDTFKD LINNLHDFNPSFLDGFDEKPLLDIKTEDGIKVEPPNAQDLINSLNVKSET
glghgfggfgvglgldpqsmkmrpgvgfqngpngnanagnggptaggggggngpgg	570-625	
	626-658	LMSEHSLAAQTLKQMAEQHQHKSAMGGMGGFHV
pphgmqqqqpqqqqqapqqqqqqhgqmmggpgggqqqqqqqqp	659-701	
	702-721	RYNDYGGGFPNDFAMGPNPT
qqqqqhlppqfhq	722-734	
	735-769	KAPGGGPGMNVQQNFLDIKQELFYSSPNDFDLKHL
qqqqamqqqqqqqqqqqqqqqhhaqqqqqhpngpnmgvpmggagnfakqqqqqvptpqqq qqqqlqqqqqqysspfsnqnananflncpprggpqgnqapgnmpqqqqqqpqqqqqpprgp qsnpnavpggnaanatqqqqqqqqqqqqqqqqqqqqqqqqatttttlqmkqtqqlhisqqqg gshgiq	770-955	
	956-976	VSAGQHLHLSSDMKSNVSVAA
qqgvffsqqqaaqqqqqqqqqpgnagpnpqqqqqqphggnaganggqpngpqqqqpnqnm nnsnvpsdgfslsqsqsmnftqqqqqqaaaaaaaaaaqqqqaaaaqqqqqqvppnmrqr qtqaqaaaaaaaaaaagaqaaananggpggnvplmqqqqqtpggvpvgagsg	977-1148	
	1149-1169	NASVGVPVSAGGFNNGAMNQL
ggpmggmpgmqmggpggvpinpmqmnpnggapnaqmmmggnggqpvpaasqakflqqqqi mraqamqhqqqvqqhmagarppppeynatkaqlmqaqmmqqtvgggggggvgvgvgvggg vggggagrfpnsaaqaaamrrmtqq	1170-1315	
	1316-1336	PIPPSGPMMRPQHAAMYMQQH
ggagggprggmggpyggggvggaggpmgggggggqqqqrppnvqvtpdgmpmgsqqewrh mmmtqqqqqmgfgpggpmrqgpggfnggnfmpngapnapgngpnggggggmmpgpngpqm qltpaqmqqqhmrqqqqqqhmgpgggggggggnmqmqqllqqqqnaaaggggqmmatqmq mtsihmsqtqqqqqltmqqqqfvqststttthqqqqqlqlqmqsqsggpggngpsnnnga nqaggvgvgvgvgvgvgvvgssatiasassisqtins	1337-1613	
	1614-1655	VVANSNDLCLEFLDNLPDGNFSTQDLINSLDNDNFNIQDILQ

(b)

MLL-AF4 FUSION PROTEIN

LOW-COMPLEXITY SEGMENTS		HIGH-COMPLEXITY SEGMENTS
mahscrwrfparpgttgggggggrrglgggprqrvpalllppgppvggggpgappsppav aaaaaaagssgagvpggaaaasaaasssaassssssssssassgpa	1-104	
	105-133	LLRVGPGFDAALQVSAAIGTNLRRFRAVF
gesgggggsge	134-144	
	145-203	LTTQIPCSWRTKGHIHDKKTEPFRLLAWSWCLNDEQFLGFGSDEEVRVRSPTRSPSVKT
sprkprgrprsgsdrnsailsdpsvfsplnksetksgdkikkkdsksiekk	204-254	
	255-487	RGRPPTFPGVKIKITHGKDISELPKGNKEDSLKKIKRTPSATFQQATKIKKLRAGKLSPL KSKFKTGKLQIGRKGVQIVRRRGRPPSTERIKTPSGLLINSELEKPQKVRKDKEGTPPLT KEDKTVVRQSPRRIKPVRIIPSSKRTDATIAKQLLQRAKKGAQKKIEKEAAQLQGRKVKT QVKNIRQFIMPVVSAISSRIIKTPRRFIEDEDYDPPIKIARLESTPNSRFSAP
scgssekssaasqhssqmssdssrsssspsvdtstds	488-523	
	524-555	QASEEIQVLPEERSDTPEVHPPLPISQSPENE
sndrrsrrysvsersfgsrttkklstlqsapqqqtssspppplltppplqpass	556-610	
	611-702	ISDHTPWLMPPTIPFGLCSNNPLTSPFLPASTAPMQGKRKSILREPTFRWTSLKHSRSEP QYFSSAKYAKEGLIRKPIFDNFRPPPLTPEDV
gfasgfsasgtaasa	703-717	
	718-759	RLFSPLHSGTRFDMHKRSPLLRAPRFTPSEAHSRIFESVTLP
snrtsagtsssgvsnr	760-775	
	776-802	KRKRKVPSPIRSEPRSPSHSMRTRSGR
lsssalspltppssvsssls isvspl	803-828	
	829-861	ATSALNPTFTFPSHSLTQSGESAEKNQRPRKQT
sapaepfsssaptplfp	862-878	
	879-909	WFTPGSQTERGRNKDKAPEELSKDRDADKSV
ekdksrerdrerekenkresrkekrkk	910-936	
	937-962	GSEIQSSSALYPVGRVSKEKVVGEDV
atsssakkat	963-972	
	973-1032	GRKKSSSHDSGTDITSVTLGDTTAVKTKILIKKGRGNLEKTNLDLGPTAPSLEKEKTLCL
stpsstvkhstss	1033-1046	
	1047-1225	IGSMLAQADKLPMTDKRVASLLKKAKAQLCKIEKSKSLKQTDQPKAQGQESDSSETSVRG PRIKHVCRRAAVALGRKRAVFPDDMPTLSALPWEEREKILFSMGNDDKSSIAGSEDAEPL APPIKPIKPVTRNKAPQEPPVKKGRRSRRCGQCPGCQVPEDCGVCTNCLDKPKFGGRNI
kkqcckmrkcqniqwmpskaylqkqakavkkkekksktsekkdskessvvknvvdssqkp tpsaredpapkkssseppprkpveek	1226-1311	
	1312-1346	SEEGNVSAPGPESKQATTPASRKSSKQVSQPALVI
ppqpttqpprkevpkttpsepkkkqppppesgpeqskqkkvaprpsipvkqkpkekekp ppvnkq	1347-1412	
	1413-1523	ENAGTLNIFSTLSNGNSSKQKIPADGVHRIRVDFKQTYSNEVHCVEEILKEMTHSWPPPL TAIHTPSTAEPSKFPFPTKDSQHVSSVTQNQKQYDTSSKTHSNSQQGTSSM
leddlqlsdsedsdseqtpekppsssappsapqslpepvasahsssaesestsdsdsssd sesesssdseeneplstpapepepptt	1524-1611	
	1612-1659	NKWQLDNWLTKVSSQLRHQRAPGAQSPHGGTQRVRAAATVPRVRSILN
pkilplkapakpprppeaphpgkrscqkspaqqeppqrq	1660-1698	
	1699-1765	TVGTKQPKKPVKASARAGSRTSLQGEREPGLLPYGSRDQTSKDKPKVKTKGRPRAAASNE PKPAVPP
ssekkkhkss	1766-1775	
	1776-1808	LPAPSKALSGPEPAKDNVEDRTPEHFALVPLTE
sqgpphsgassrtsg	1809-1823	
	1824-1894	CRQAVVVQEDSRKDRLPLPLRDTKLLSPLRDTPPPQSLMVKITLDLLSRIPQPPGKGSRQ RKAEDKQPPAG
kkhssekrsdsssklakkrkgeaerdcdnkkirlekeiksqsssssshkessktkpsr psqsskkemlppppvsssqkpakpa	1895-1981	

Fig. 2. Automatic segmentation of amino acid sequences into segments of low-complexity and high-complexity by the SEG algorithm [8•]. The proteins are **(a)** *Drosophila* virilis, translated product of the *mastermind* gene ([14•]; GenBank accession number M92914, locus name DROMASTM); **(b)** the MLL-AF4 translocation fusion protein in a human acute leukemia ([15•]; GenBank accession number L22179, locus name HUMMLLAF4F). All the lines of the sequence read from left to right and the order of the segments in each polypeptide runs from top to bottom, as shown by the central columns of residue numbers. The stringency of segmentation was the same for both proteins.

Return to Figure reference in text 1

Compositional complexity is a function of the compositional state of a sequence segment or window. For a 12 residue peptide window, for example, the ordered vector of numbers (3,3,2,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0), representing counts for the various amino acids, describes one of the 77 possible complexity states. Many possible sequences and amino acid compositions, with different residue types corresponding to the 20 numbers, share this complexity state. The local compositional complexity (K_1) of a window of length L is defined as:

$$K_1 = \frac{1}{L} \log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right)$$

where n_i is the 20 numbers in the complexity state vector. K_1 measures the information per position needed, given the window's composition, to specify a particular residue order, regardless of which amino acids correspond to the 20 numbers in the vector and independently of the actual probabilities for appearance of the various amino acids. The probability (P_0) for the occurrence of a given complexity state, based on this assumption of uniform probabilities for the 20 amino acids, is:

$$P_0 = \frac{1}{20^L} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right) \left(\frac{20!}{\prod_{k=0}^{L} r_k!} \right)$$

where r_k is the number of times that the number k occurs among n_i.

The SEG algorithm [8•][11] further develops this theory in the spirit of unbiased exploration of sequence data. The sequence or database is initially treated as a heterogeneous mixture of regions with unknown statistical properties; attempts may then be made to infer these properties. SEG first identifies all low-complexity segments in a sequence or database at a defined level of stringency, then local subsequences of minimal P_0 of any length are reported. These rigorously optimized segments correspond well with intuitive concepts of simple sequences, as illustrated below.

Masking low-complexity sequences for database searches

Low-complexity sequences create a serious problem for database search algorithms based on pair-wise sequence comparison and alignment, because they are not encompassed by the random model used by these methods to evaluate local alignment statistics [12•]. Algorithms such as SEG may be used to mask automatically the low-complexity segments in database search query sequences, replacing their residues with 'x' characters. This removes the potential confusion caused by overwhelmingly large output lists, with many spuriously high scores due to compositional bias, in which interesting similarities may be inconspicuously buried [8•][12•][13]. An important recent application of such automated masking has been in the production of precomputed sets of all amino acid sequence homologues ('neighbors') released with the NCBI databases and incorporated into the ENTREZ retrieval system [3]. Since September 1992, the entire sequence databases and all updates have been masked before searches, in order to avoid spurious matches in the neighbor lists.

The abundance of low complexity in protein sequence databases

Fig. 1 illustrates some of the recent results obtained from complexity measures on sequence databases and their subsets. Details of calculations of this type and the cleaned-up databases used are described in [8•][11]. The distribution of complexity K_1 shows a remarkable contrast between the sequence database SWISS-PROT (26 August 1993 release; [2]), the sequences in crystal structures from PDB (July 1993 release; [4]) and random sequences (generated with SWISS-PROT amino acid frequencies). Compared with random sequences, SWISS-PROT sequences display a substantial excess of low-complexity windows, amounting to approximately 25% of the residues. In contrast, the much smaller set of PDB sequences is very close to random in its distribution of complexity, less than 1% of the residues falling outside the random range. In species-specific subsets of SWISS-PROT, all groups of organisms (bacteria, protists, yeast, fungi, plants and animals) show an abundance of low-complexity segments, varying from 11% (*Escherichia coli*) through 24% (human) to 36% (*Drosophila*) of the residues [11]. Fig. 1 shows *Drosophila* results as an example.

The upper end of 'low complexity' includes some relatively well understood, long, non-globular domains of proteins, in addition to many new classes [11]. These include collagens (Fig. 1), which exhibit a bimodal distribution of local complexity corresponding to the windows from the triple-helical rod and globular domains, and coiled-coil proteins such as myosins (Fig. 1) for which the SEG algorithm may be used (with appropriate parameters) to separate cleanly the globular head sequences from the variably repeated coiled-coil sequences. Similar distributions are found with keratins, other intermediate filament proteins, proteoglycan core proteins, mucins, elastins and fibrins [11]. These results support the concept of a general correlation between non- random (reduced) compositional complexity and non-globular elongated structure. Any tendency, however slight, towards regular sequence repetition, as occurs in these helical structures, will inevitably generate relatively lower values of K_1 at sufficiently long window length.

If we discount these known non-globular, 'medium-complexity' structures, a conservative estimate still counts 15% of the residues in the sequence database as occurring in segments of strikingly low complexity [8•][11]. The dramatic meaning of 'strikingly' is illustrated by two important examples published in 1993 (Fig. 2; [14•][15•]). The SEG algorithm also shows that approximately 34% of the sequence entries in SWISS-PROT contain at least one such (relatively stringently defined) low-complexity segment; this increases to 56% if stringency is relaxed to include potential long non-globular regions of non-random low complexity [11].

Speculations that extrapolate from the current sequence and structure databases of natural proteins suggest that there may be only a limited number of 'ancient conserved regions' or alignable sequence families, corresponding to globular topological folds [16][17]. Possibly, extrapolating even further into uncertainty, globular domains that are conserved across different phyla may account for little more than half of the residues in the total set of genome-encoded protein products, with low-complexity segments, non-globular regions, and relatively non-conserved globular domains making up the rest.

Randomness, compactness and structural uniqueness

Evidently, the PDB almost exclusively contains sequences of high compositional complexity, similar to random sequences, presumably because compact globular proteins that form relatively unique folds and highly-ordered crystals have this property (Fig. 1; [11]). Thus, the 'globular perspective' on the current database of protein sequences is to think of quasi-random, high compositional complexity as the norm, with low-complexity sequences deviating from this norm to different extents. There is a long history of the idea that globular structures resemble random sequences in local composition, although interesting subtle differences exist (Finkelstein, this issue, pp 422-428; [18][19]). This idea is now supported by these rigorous complexity measurements (Fig. 1; [4][11]) and other database analyses [20][21]. Conversely, random permutations of the sequences of globular proteins may have a high probability of forming compact folded structures [18]. Perhaps, as a general design principle for proteins and protein-like polymers, high compositional complexity is a necessary condition for the formation of a unique globular fold that has intricate, tight packing of side chains into a solvent-excluding core. Low-complexity sequences, even those with an appropriate proportion of hydrophobic side chains to favour relatively compact structures, are more likely to generate relatively extended coiled or helical structures or mobile states resembling molten globules.

Structures and dynamics

Low-complexity sequences are expected to show a wide range of deviations from structural compactness and conformational uniqueness, from helical ellipsoidal domains, through elongated rods, to mobile coils resembling synthetic homopolymers or co-polymers. Recent insights into this difficult area have come from several interesting crystal structures of proteins with well defined, elongated, helical, non-globular domains [22][23•][24•][25•][26]. These domains show sequence repeats some of which are very subtle. The SEG algorithm can distinguish the globular and non-globular domains of these proteins on criteria of compositional complexity (JC Wootton, unpublished data). These non-globular architectures have packed, hydrophobic interiors that are protected from solvent exposure to different extents, but these contrast with typical globular proteins in being narrow and elongated, with stacked patterns of interlocked side chains. These domains are essentially intermediate between globular structures and long rods such as alpha-helical coiled-coils and collagen, in which interlocking of hydrophobic side chains occurs to a limited extent in a relatively solvent-exposed environment.

Table 1. Interactions, functions and phenotypes recently attributed to low-complexity segments.

Protein(s)	Function	Example of low-
Involved in tumorigenesis or tumor suppression		
MLL[a] fusion proteins in acute lymphoblastic leukemia	Ser-Pro-rich segments fused to MLL by chromosomal translocations t(4:11) and t(11:9) involved in tumorigenesis [15•][33•]	PPSSSAPPSAPQSLPEPVASAHSSSAESESTSDSDSSSDSESI
Tumor suppressor protein WT1	Transcriptional regulation in Wilms' tumor acting independently of DNA- binding domain [34]	SLGGGGGCALPVSGAAQWAPVLDFAPPGASAYGSLGGP/
Interactions with RNP complexes		
SC35, SF2/ASF, U2AF, Tra-2 (SR proteins)	Activating interactions in commitment to splicing of pre-mRNA [38]	GRRSRSPRRRRRSRSRSRSRSRSRSRSRYSRSKSRSRTRSRS
Nucleolin, fibrillarin, hnRNP core proteins	Interactions in formation of pre-mRNA processing complexes, possibly RNA helix destabilizing. Arginines variably dimethylated[39•][40]	GGRGGGRGGFGGRGGGRGGRGGFGGRGRGGFGGRGGFI
hnRNP M proteins	Pre-mRNA binding [36]	RMGPGIDRLGGAGMERMGAGLGHGMDRVGSEIERMGL\
Direct DNA or RNA binding demonstrated		
H1 histones and many other DNA-binding proteins	'SPKK' motif; minor groove DNA binding NMR study and model building [41]	SPKKSPRK
methionyl-tRNA synthetase	RNA-binding helix-loop: sequence simplified with engineered substitutions to Ala and Ser [37•]	AAVSAIAALASAANRYVSESSPWAVAKSEA
HIV Tat and Rev proteins	Binding to TAR RNA stem-loop [42]	RKKRRQRRRPPQNS
Interactions in transcription and transcriptional regulation		

RNA polymerase II, largest subunit	Repeats of carboxy-terminal domain variably phosphorylatedon Ser/Thr; multiple interactions in transcription postulated[43]	YSPTSPS and variants repeated 26-52 times
Transcription factors Sp1, CREB	Activates transcription by binding to TFIID 110KDa subuniT (TAF110), hydrophobic residues rather than glutamines implicated [44•]	GPNGQVSWQTLQLQNLQVQNPQAQ
Fushi tarazu homeotic protein	Forms gene-specific transcription activation complexes by binding to TFIIB [45]	GYTAMLPPLEATSTATTGAPSVPVPMYHHHQTTAAYPAY
Transcription factor HNF-1A	Liver-specific transcription activation [46]	QAQSVPVINSMGSSLTTLQPVQFSQPLHPSYQQPLMPPVQ
Other protein-protein interactions		
3BP1, 3BP2 (GAP-Rho-like)	Specific binding to SH3 domains of many cytoskeletal and signalling proteins [47]	PPAYPPPPVP
GroES	Mobile loop; conformational adaptability in intersubunit interactions [48]	KRKEVETKSAGGIVLTGSAA
Linker or extended structural or dynamic role		
TonB protein	Bridges periplasm in gram-negative bacteria, not essential for energy transduction [49•]	PPQAVQPPPEPVVEPEPEPEPIPEPPKEAPVVIEKPKPKPKP
Tyrosine kinase receptor in breast carcinoma cells	Long interdomain-linker sequences flanking transmembrane segment [50]	NNSSPALGGTFPPAPWWPPGPPPTNFSSLELEPRGQQPVA

Pyruvate dehydrogenase and many multidomain proteins	Pro-Ala-rich interdomain linkers [51]	APAAAPAKQEAAAPAPAAKAEAPAAAPAAK
Many bacterial proteins	’Q-linker’ interdomain linkers [52]	QQRQQQEQGRLDRLKQQMTAGKL
Extended α-helices		
Type-I antifreeze polypeptide	Binds to ice crystals and prevents growth [53]	TASDAAAAAALTAANAKAAAELTAANAAAAAAATAR
Caldesmon	Single solvent-exposed charged extended helix [54]	LKAEEEKKAAEEKQKAEEEKKAAEERERAKAEEEKRAAI
Synthetic peptides, polymerized on gold surfaces	Aligned dipoles for optical switching (molecular electronics) [55]	Polyalanine, polyphenylalanine
Synthetic peptides	Variants for helix stability studies [56]	ADAAARDAAARDAAARY

[a]MLL, myeloid-lymphoid leukemia. [b]The examples of sequences are shown for illustration only: the origi details.

Return to table reference 1 , 2

In contrast, very little is known about the structures and dynamics of the abundant low-complexity regions interspersed within the sequences of complex multidomain proteins. Many of these are hydrophilic and likely to be relatively mobile. Some may form single solvent-exposed alpha helices, as suggested by i+3 and i+4 complementary charge patterns, but most lack such patterns. Their conformational states are likely to be determined by a number of factors: torsional constraints in the polypeptide backbone; electrostatic interactions and hydrogen bonds involving solvent molecules and ions; and hydrodynamic effects. Multiple chains may interact, possibly with cross-linking in some cases, and complex processive assembly mechanisms may occur. Conformational adaptability in interactions with intracellular and extracellular molecular complexes may be crucial for cellular mechanics, dynamics and morphogenesis. This is an important area for future research.

Evolution and functions

Many DNA sequences encoding low-complexity regions of proteins evidently evolve rapidly by processes such as recombinational repeat expansion and deletion, replication slippage and a high frequency of substitution mutations. In this respect, such sequences resemble other simple DNA sequences found in non-coding sequences such as variable number tandem repeats and microsatellites. Polymorphic trinucleotide repeats implicated in several human genetic diseases [27] provide extreme examples. Some of these repeats occur in coding sequences, generating variable-length homopolymers, such as polyglutamine, whereas others are in untranslated regions.

Rapid evolution involving length changes has been clearly demonstrated recently by several interspecific comparisons of low-complexity sequences [14•][28][29][30][31], with interesting implications for studies of the mechanisms of mutational dynamics [10][27][32]. These studies emphasize the importance of questions about the range of phenotypic consequences of these mutational changes and the extent to which the observed spectrum of low-complexity sequence features is generated by mutational drive rather than selection at the protein level. Do rapid mutational changes in low-complexity segments generate important phenotypic innovation in the evolution of organisms with complex development and morphology? What is the magnitude of the genetic load imposed by this type of genome/phenotype flux?

Recently, several important interactions, functions and phenotypes have been attributed to low-complexity segments (Table 1). These include tumorigenesis, specific DNA and RNA binding, interactions in transcriptional regulation, selection of pre-mRNAs for splicing in nuclear RNP complexes, protein-protein interactions in signal transduction, control of protein folding and turnover, and specific roles in cellular and extracellular mechanics. Clearly, many low-complexity regions in multidomain proteins are not merely present in non-essential regions as a result of mutational drive, nor do they simply act as linkers.

Conclusion

The importance and urgency of continued research on low-complexity regions is underlined by their range of important interactions and functions (Table 1), including their involvement in human molecular diseases [15•][27][33•][34] and as epitopes in autoimmune diseases [21][35]. They are, however, characteristically difficult to study at the molecular level.

They are often heterogeneously modified, for example by phosphorylation, methylation or glycosylation [36][37•]. Many are conformationally mobile and others form defined non-globular structures which may be polymorphic or cross-linked and assembled into relatively intractable complexes. Few readily form ordered crystals. Repetitive sequences are also difficult for NMR studies because of ambiguities in unique nuclear Overhauser enhancement (NOE) assignments.

In the face of these difficulties, it is helpful to have clear statistical criteria to identify low-complexity sequences and classify them by attributes such as composition, k-gram patterns and periodicity [7][8•][9][10][11]. There is also scope for critical and imaginative theoretical modelling studies based on principles of protein conformation and architecture, although these may depend more on relatively unfamiliar considerations of torsional and hydrodynamic constraints, solvent interactions and conformational entropy, rather than on empirical potentials of mean force derived from the database of globular structures. However, the major challenges are for ingenious combinations of experimental methods involving new genetic expression systems suitable for relatively heterogeneous and biochemically intractable molecules, engineered sequence changes, detailed molecular analyses by several spectroscopic and other physical techniques, structural biology in favourable cases, and a wide range of functional and phenotypic studies at all levels.

Acknowledgements

There is a very large body of interesting literature on low- complexity sequences and relevant physicochemical studies on polymers. I sincerely apologize to all investigators whose work is not cited in this review because of limitations of scope or space.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as.

● of special interest.

●● of outstanding interest.

1. Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A: **The PIR-International Databases.** *Nucleic Acids Res* 1993, **21 :** 3089-3092.
   [Entrez], Return to reference citation 1

2. Bairoch A, Boeckmann B: **The SWISS-PROT Protein Sequence Databank, Recent Developments.** *Nucleic Acids Res* 1993, **21 :** 3093- 3096.
   [Entrez], Return to reference citation 1 , 2 , 3

3. Benson D, Lipman DJ, Ostell J: **GenBank.** *Nucleic Acids Res* 1993, **21 :** 2963-2965.
   Return to reference citation 1 , 2

4. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures.** *J Mol Biol* 1977, **112 :** 535-542.
   Return to reference citation 1 , 2 , 3 , 4

5. Karlin S, Brendel V: **Charge Configurations in Viral Proteins.** *Proc Natl Acad Sci U S A* 1988, **85 :** 9396-9400.
   [Entrez], Return to reference citation 1

6. Brendel V, Karlin S: **Association of Charge Clusters with Functional Domains of Cellular Transcription Factors.** *Proc Natl Acad Sci U S A* 1989, **86 :** 5696-5702.
   Return to reference citation 1

7. Karlin S, Brendel V: **Charge and Statistical Significance in Protein and DNA Sequence Analysis.** *Science* 1992, **257 :** 39-49.
   Return to reference citation 1 , 2

8. ● Wootton JC, Federhen S: **Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases.** *Comput Chem* 1993, **17 :** 149-163.
   The first use of rigorous compositional complexity measures, of which several are compared, to analyze protein sequences and databases. Introduces the SEG algorithm for partitioning sequences

and databases according to local compositional complexity and defining optimized low-complexity segments at different levels of stringency.
Return to reference citation 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 , 10

9. Salamon P, Konopka AK: **A Maximum Entropy Principle for Distribution of Local Complexity in Naturally Occurring Nucleotide Sequences.** *Comput Chem* 1992, **16 :** 117-124.
Return to reference citation 1 , 2 , 3

10. Salamon P, Wootton JC, Konopka AK, Hansen LK: **On the Robustness of Maximum Entropy Relationships for Complexity Distributions of Nucleotide Sequences.** *Comput Chem* 1993, **17 :** 135- 148.
Return to reference citation 1 , 2 , 3 , 4

11. Wootton JC: **Non-Globular Domains in Protein Sequences: Automated Segmentation Using Complexity Measures.** *Comput Chem* 1994, **18 :** in press.
[Entrez], Return to reference citation 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 , 10 , 11 , 12 , 13

12. • Altschul SF, Boguski M, Gish W, Wootton JC: **Issues in Searching Molecular Sequence Databases.** *Nature Genet* 1994, **6 :** 119-129.
This paper discusses issues such as sequence redundancy, repetitiveness and low-complexity, that seriously confound the sequence comparison algorithms commonly used for database searching. It also describes how automated masking methods may be applied to overcome these problems and facilitate the discovery of biologically significant sequence similarities.
[Entrez], Return to reference citation 1 , 2

13. Claverie JM, States DJ: **Information Enhancement Methods for Large Scale Sequence Analysis.** *Comput Chem* 1993, **17 :** 191-201.
Return to reference citation 1

14. • Newfeld SJ, Schmid AT, Yedvobnick B: **Homopolymer Length Variation in the *Drosophila*Gene *mastermind*.** *J Mol Evol* 1993, **37 :** 483-495.
The mastermind gene controls a crucial switch in neurogenesis in Drosophila embryos. The translated amino acid sequence consists almost entirely of homopolymers and other low-complexity segments (Fig. 1). This paper compares the sequences from two Drosophila species and begins to address interesting questions of the balance between mutational drive and selection at the protein level as determinants of the evolution of low-complexity sequences.
[Entrez], Return to reference citation 1 , 2 , 3

15. • Nakamura T, Alder H, Gu Y, Prasad R, Canaani O, Kamada M, Gale RP, Lange B, Crist WM, Nowell PC, Croce CM, Canaani E: **Genes on Chromosomes 4, 9, and 19 Involved in 11q23 Abnormalities in Acute Leukemia Share Sequence Homology and/or Common Motifs.** *Proc Natl Acad Sci U S A* 1993, **90 :** 4631-4635.
This paper and [33•] establish the involvement of a set of similar extensive Ser-Pro-rich domains, fused to the MLL gene on chromosome 11 by translocations, in a wide range of acute myeloid-lymphoid leukemias.
[Entrez], Return to reference citation 1 , 2 , 3 , 4 , 5

16. Green P, Lipman D, Hillier L, Waterston M, States DJ, Claverie JM: **Ancient Conserved Regions in New Gene Sequences.** *Science* 1993, **259 :** 1711-1716.
[Entrez], Return to reference citation 1

17. Chothia C: **Proteins. One Thousand Families for the Molecular Biologist.** *Nature* 1992, **357 :** 453-544.
    Return to reference citation 1

18. Ptitsyn OB: **Random Sequences and Protein Folding.** *J Mol Struct* 1985, **123 :** 45-65.
    Return to reference citation 1 , 2

19. Chothia C, Finkelstein AV: **The Classification and Origins of Protein Folding Patterns.** *Ann Rev Biochem* 1990, **59 :** 1007-1039.
    [Entrez], Return to reference citation 1

20. Rao S, Zhu QL, Vajda S, Smith T: **The Local Information Content of the Protein Structural Database.** *FEBS Lett* 1993, **322 :** 143-146.
    [Entrez], Return to reference citation 1

21. Dohlman JG, Lupas A, Carson M: **Long Charge-Rich Alpha- Helices in Systemic Autoantigens.** *Biochem Biophys Res Commun* 1993, **195 :** 686-696.
    [Entrez], Return to reference citation 1 , 2

22. Ellenberger TE, Brandl CJ, Struhl K, Harrison SC: **The GCN4 Basic Region Leucine Zipper Binds DNA as a Dimer of Uninterrupted Alpha Helices: Crystal Structure of the Protein-DNA Complex.** *Cell* 1992, **71 :** 1223-1237.
    [Entrez], Return to reference citation 1

23. • Yoder MD, Keen NT, Jurnak F: **New Domain Motif: The Structure of Pectate Lyase C, a Secreted Plant Virulence Factor.** *Science* 1993, **260 :** 1053-1057.
    This paper and [24•][25•] describe crystal structures of proteins with novel, elongated, helical domain architectures: a parallel beta helix, a parallel alpha-beta helix and a parallel beta-roll. These contrast strongly with the globular (spheroidal or ellipsoidal) domains of almost all other protein crystal structures. The helical structures reflect the presence of repeats in the amino acid sequences, although these are very subtle in the case of pectate lyase C.
    Return to reference citation 1 , 2 , 3

24. • Kobe B, Deisenhofer J: **Crystal Structure of Porcine Ribonuclease Inhibitor, a Protein with Leucine-Rich Repeats.** *Nature* 1993, **366 :** 751-756.
    See [23•].
    [Entrez], Return to reference citation 1 , 2

25. • Baumann U, Wu S, Flaherty KM, McKay DB: **Three-Dimensional Structure of the Alkaline Protease of *Pseudomonas aeruginosa*: A Two-Domain Protein with a Calcium Binding Parallel Beta Roll Motif.** *EMBO J* 1993, **12 :** 3357-3364.
    See [23•].
    [Entrez], Return to reference citation 1 , 2

26. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, Branton D: **Crystal Structure of the Repetitive Segments of Spectrin.** *Science* 1993, **262 :** 2027-2030.
    [Entrez], Return to reference citation 1

27. Riggins GJ, Lokey LK, Chastain JL, Leiner HA, Sherman SL, Wilkinson KD, Warren ST: **Human**

Genes Containing Polymorphic Trinucleotide Repeats. *Nature Genet* 1992, **2 :** 186-191.
[Entrez], Return to reference citation 1 , 2 , 3

28. Schlotterer C, Tautz D: **Slippage Synthesis of Simple Sequence DNA.** *Nucleic Acids Res* 1992, **20 :** 211-215.
[Entrez], Return to reference citation 1

29. Hancock JM: **Evolution of Sequence Repetition and Gene Duplications in the TATA-Binding Protein TBP (TFIID).** *Nucleic Acids Res* 1993, **21 :** 2823-2830.
[Entrez], Return to reference citation 1

30. Whitfield LS, Lovell-Badge R, Goodfellow PN: **Rapid Sequence Evolution of the Mammalian Sex-Determining Gene SRY.** *Nature* 1993, **364 :** 713-715.
[Entrez], Return to reference citation 1

31. Tucker PK, Lundrigan BL: **Rapid Evolution of the Sex-Determining Locus in Old World Mice and Rats.** *Nature* 1993, **364 :** 715-717.
[Entrez], Return to reference citation 1

32. White SH, Jacobs RE: **The Evolution of Proteins from Random Amino Acid Sequences. I. Evidence from the Lengthwise Distribution of Amino Acids in Modern Protein Sequences.** *J Mol Evol* 1993, **36 :** 79-95.
[Entrez], Return to reference citation 1

33. • Corral J, Forster A, Thompson S, Lampert F, Kaneko Y, Slater R, Kroes WG, Van Der Schoot CE, Ludwig WD, Karpas A, et al: **Acute Leukemias of Different Lineages Have Similar MLL Gene Fusions Encoding Related Chimeric Proteins Resulting from Chromosomal Translocation.** *Proc Natl Acad Sci U S A* 1993, **90 :** 8538-8542.
This paper and [15•] establish the involvement of a set of similar extensive Ser-Pro-rich domains, fused to the MLL gene on chromosome 11 by translocations, in a wide range of acute myeloid-lymphoid leukemias.
[Entrez], Return to reference citation 1 , 2 , 3

34. Madden SL, Cook DM, Rauscher FJ: **A Structure-Function Analysis of Transcriptional Repression Mediated by the WT1, Wilms' Tumor Suppressor Protein.** *Oncogene* 1993, **8 :** 1713-1720.
[Entrez], Return to reference citation 1 , 2

35. Brendel V, Dohlman J, Blaisdell BE, Karlin S: **Very Long Charge Runs in Systemic Lupus Erythematosus-Associated Autoantigens.** *Proc Natl Acad Sci U S A* 1991, **88 :** 1536-1540.
[Entrez], Return to reference citation 1

36. Datar KV, Dreyfuss G, Swanson MS: **The Human hnRNP M Proteins: Identification of a Methionine/Arginine-Rich Repeat Motif in Ribonucleoproteins.** *Nucleic Acids Res* 1993, **21 :** 439-446.
[Entrez], Return to reference citation 1 , 2

37. • Kim S, Ribas de Pouplana L, Schimmel P: **Diversified Sequences of Peptide Epitope for Same-RNA Recognition.** *Proc Natl Acad Sci U S A* 1993, **90 :** 10046-10050.
This paper reports interesting sequence simplifications engineered into the essential RNA binding

helix-loop of methionyl-tRNA synthetase. Variants with up to 17 alanine and serine residues (out of 30) retain activity and specificity.
[Entrez], Return to reference citation 1 , 2

38. Fu XD: **Specific Commitment of Different Pre-mRNAs to Splicing by Single SR Proteins.** *Nature* 1993, **365 :** 82-85.
[Entrez], Return to reference citation 1

39. • Najbauer J, Johnson BA, Young AL, Aswad DW: **Peptides with Sequences Similar to Glycine, Arginine-Rich Motifs in Proteins Interacting with RNA Are Efficiently Recognized by Methyltransferase(s) Modifying Arginine in Numerous Proteins.** *J Biol Chem* 1993, **268 :** 10501-10509.
Arginine-specific methyltransferases demonstrate strong binding and activity for Gly-Arg-rich sequences resembling those of nucleolin, fibrillarin and hnRNP core proteins. This provides a clear demonstration of a specific interaction involving low- complexity segments.
[Entrez], Return to reference citation 1

40. Ghisolfi L, Joseph G, Amalric F, Erard M: **The Glycine- Rich Domain of Nucleolin Has an Unusual Supersecondary Structure Responsible for Its RNA-Helix-Destabilizing Properties.** *J Biol Chem* 1992, **267 :** 2955-2959.
[Entrez], Return to reference citation 1

41. Suzuki M, Gerstein M, Johnson T: **An NMR Study on the DNA- Binding SPKK Motif and a Model for Its Interaction with DNA.** *Protein Eng* **6 :** 565-574.
[Entrez], Return to reference citation 1

42. Delling U, Roy S, Sumner-Smith M, Barnett R, Reid L, Rosen CA, Sonenberg N: **The Number of Positively Charged Amino Acids in the Basic Domain of Tat is Critical for Trans-Activation and Complex Formation with TAR RNA.** *Proc Natl Acad Sci U S A* 1991, **88 :** 6234-6238.
[Entrez], Return to reference citation 1

43. Serizawa H, Conaway JW, Conaway RC: **Phosphorylation of C Transcription.** *Nature* 1993, **363 :** 371-374.
[Entrez], Return to reference citation 1

44. • Gill G, Pascal E, Tseng ZH, Tjian R: **A Glutamine-Rich Hydrophobic Patch in Transcription Factor Sp1 Contacts the dTAFII110 Component of the *Drosophila*TFIID Complex and Mediates Transcriptional Activation.** *Proc Natl Acad Sci U S A* 1994, **91 :** 192-196.
Mutagenesis experiments implicate hydrophobic residues within the glutamine-rich activation domain in a crucial specific interaction in transcription activation. The glutamine residues themselves are evidently not critical. This clarifies the distinction between glutamine richness per se , the molecular significance of which is still poorly understood, and the critical binding specificities.
[Entrez], Return to reference citation 1

45. Colgan J, Wampler S, Manley JL: **Interaction Between a Transcriptional Activator and Transcription Factor TFIIB *in vivo*.** *Nature* 1993, **362 :** 549-553.
[Entrez], Return to reference citation 1

46. Toniatti C, Monaci P, Nicosia A, Cortese R, Ciliberto G: **A Bipartite Activation Domain is Responsible for the Activity of Transcription Factor HNF1/LFB1 in Cells of hepatic and**

Non-Hepatic Origin. *DNA Cell Biol* 1993, **12 :** 199-208.
[Entrez], Return to reference citation 1

47. Ren R, Mayer BJ, Cicchetti P, Baltimore D: **Identification of a Ten-Amino Acid Proline-Rich SH3 Binding Site.** *Science* 1993, **259 :** 1157-1161.
[Entrez], Return to reference citation 1

48. Landry SJ, Zeilstra-Ryalls J, Fayet O, Georgopoulos C, Glerasch LM: **Characterization of a Functionally Important Mobile Domain of GroES.** *Nature* 1993, **364 :** 255-258.
[Entrez], Return to reference citation 1

49. Larsen RA, Wood GE, Postle K: **The Conserved Proline-Rich Motif is not Essential for Energy Transduction by *Escherichia coli*TonB Protein.** *Mol Microbiol* 1993, **10 :** 943-953.
[Entrez], Return to reference citation 1

50. Johnson JD, Edman JC, Rutter WJ: **A Receptor tyrosine Kinase Found in Breast Carcinoma Cells Has an Extracellular Discoidin I-Like Domain.** *Proc Natl Acad Sci U S A* 1993, **90 :** 5677-5681.
[Entrez], Return to reference citation 1

51. Turner SL, Russell GC, Williamson MP, Guest JR: **Restructuring an Interdomain Linker in the Dihydrolipoamide Acetyltransferase Component of the Pyruvate Dehydrogenase Complex of *Escherichia coli*.** *Protein Eng* 1993, **6 :** 101-108.
[Entrez], Return to reference citation 1

52. Wootton JC, Drummond MH: **The Q-linker: a Class of Interdomain Sequences Found in Bacterial Multidomain Regulatory Proteins.** *Protein Eng* 1989, **2 :** 535-543.
[Entrez], Return to reference citation 1

53. Wen D, Laursen RA: **Structure-Function Relationships in an Antifreeze Polypeptide. The Role of Charged Amino Acids.** *J Biol Chem* 1993, **268 :** 16396-16400.
[Entrez], Return to reference citation 1

54. Wang CLA, Chalovich JM, Graceffa P, Lu RC, Mabuchi K, Stafford WF: **A Long Helix from the Central Region of Smooth Muscle Caldesmon.** *J Biol Chem* 1991, **266 :** 13958-13963.
[Entrez], Return to reference citation 1

55. Whitesell JK, Chang HK: **Directionally Aligned Helical Peptides on Surfaces.** *Science* 1993, **261 :** 73-76.
Return to reference citation 1

56. Huyghues-Despointes BMP, Scholz JM, Baldwin RL: **Helical Peptides with Three Pairs of Asp-Arg and Glu-Arg Residues in Different Orientations and Spacings.** *Protein Sci* 1993, **2 :** 80-85.
[Entrez], Return to reference citation 1

Abbreviations

NOE –nuclear Overhauser enhancement.

Return to Title

Author Contacts

JC Wootton, National Center for Biotechnology Information, 8th Floor, Building 38A, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Return to author list

Publication Details

Publisher: Current Biology Ltd
London

Address: Middlesex House 34-42 Cleveland Street
London
W1P 6LB
United Kingdom
Phone: 0171 580 8377
Fax: 0171 580 8428
Email: cb@cursci.co.uk

Article Information

ISSN: 0959-440X **Electronic Identifier:** 0959440X00400413
Volume Id.: 4 **Issue:** 3 **Cover Date:** June 1994
Number of figures : 2 **Number of tables :** 1 **Number of references :** 56 **Number of pages :** 9

Copyright