Proc. Indian Natl. Sci. Acad. **B71** No. 1 pp 9-18

Compositional Analysis of Proteins: An Old Theme Revisited

SRINIVASAN RAMACHANDRAN
G.N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, CSIR, Mall Road, Delhi 110 007

(Received on 12 August 2004;Accepted on 2 September 2005)

Compositional analysis of biomolecules is an age-old approach. In the years following the development of technologies for sequencing proteins and DNA, compositional analysis was largely ignored because more powerful algorithms based on sequence alignment could provide nearly direct clues on their function. The advent of genomics era has flooded the databases with numerous new sequences. Examination of the emerging voluminous sequence data has revealed that the so-called powerful sequence search algorithms have inherent limitation in uncovering the functional roles of about 40% of identified proteins. As a result compositional analysis has now returned to the forefront of bioinformatics approaches to the characterization of proteins and DNA. I have summarized recent developments in this area. A blend of old and new approaches is now required to answer difficult questions in biology.

Key Words: Compositional Analysis, Nural network, Simple sequence, Structure-function

Introduction

Proteins are the workhorses of a cell. Max Perutz referred to the living cell as a symphony orchestra without a conductor: proteins are its performing instruments (Perutz 1992). Proteins catalyze or aid numerous biological processes. It is therefore not surprising that most functional assays and sequence analyses for biological characterization are carried out with proteins. The first step in protein sequence analysis is to identify its biological function. The standard approach taken is similarity (more precisely, homology) analysis based on sequence alignment using the algorithm BLAST (Basic Local Alignment Search Tool, Altschul et al. 1990). This procedure has formed the foundation on which most database resources have been built and numerous biological inferences made.

Although BLAST is a powerful algorithm for sequence analysis, its excessive use has caused serious concern with regard to functional classification of proteins. For example, it has now become apparent that 40% of proteins are refractory to prediction of function using BLAST. In addition, it is becoming increasingly apparent that segments of sequences with compositional bias masked by BLAST and filtered out during database searches could actually be useful to reveal additional insights on function. As a result, due to these inherent limitations, compositional analysis has now returned to the forefront of analysis of proteins.

Compositional analysis of proteins precedes sequence analysis. The first characteristic of a protein that strikes the eye is its amino acid composition. An early example of using compositional information in deducing the structure of a molecule is an elegant piece of work carried out several decades ago by Ramachandran and Kartha (1955) on collagen. Collagen has a distinct amino acid composition: it is rich in glycine, proline and hydroxyproline. Glycine constitutes one third of total amino acids. Ramachandran and Kartha proposed a triple helical structure fitting the X-ray diffraction pattern. It was realized that only glycine could be accommodated at the interface of the three helices and that glycine is at the third position. This example demonstrates the use of compositional information to reveal insights into the structure and function of a molecule.

In its elementary form, compositional analysis entails estimating and computing the amount of individual constituents of a molecule. In case of

Corresponding address: *Email: ramu@igib.res.in; ramucbt@yahoo.com, Tel: 0091-11-2766-6156, x169, Fax: 0091-11-2766-7471*

proteins, the basic physico-chemical characteristics such as its solubility, isoelectric point and molecular mass are determined through compositional analysis. In the last few decades, various investigators have contributed to the growth of this field. Based on these contributions, a formal description of compositional analysis can be synthesized: Compositional analysis of proteins involves computing the frequencies of n-peptide compositions and using this property to deduce the basic characteristics of proteins such as its structure, sub-cellular location, stability and broad functional role. Note that the case n = 1 corresponds to amino acid composition, n = 2 corresponds to dipeptide composition and so on. Generally, in compositional analysis, the specific order of n-peptides is weighted less than their content. This review is an attempt to summarize some recent developments in this field and is aimed at stimulating the invention of novel applications of compositional analysis.

A fundamental distinction must be noted between compositional analysis and sequence alignment techniques. Compositional analysis is macroscopic in character: it can assign query proteins to broad functional or structural classes. Sequence alignment on the other hand is microscopic in character: it can provide nearly direct clues regarding the functional role of a query protein. However, the technique of sequence alignment suffers from one major limitation, namely, that the partner proteins in the database must have be characterized in detail. Although sequence alignment based techniques are powerful and most often form part of the first line of softwares to be used in sequence characterization, the voluminous data generated from genome sequence projects and the drive to extract biological information contained in them has forced investigators to consider both macroscopic and microscopic approaches.

The number of proteins isolated from whole cells or from sub-cellular compartments with concurrent characterization of its structural and functional features has been steadily increasing over the past decades. This repository – well collected and organized by SWISS-PROT (Boeckmann et al. 2003) - offers an opportunity to examine for compositional patterns that can be correlated to a known property of protein such as its sub-cellular location, physico-chemical character, structural class and broad functional role.

Composition as a Molecular Driver

Proteins with Distinct Compositional Characteristics

Sub-cellular location

Nishikawa and co-workers made an early attempt in this direction about 20 years ago. Their observations suggested that amino acid composition could be used to identify the overall physico-chemical characteristics of a protein such as acidic, basic or hydrophobic property. More interestingly, compositional features correlated well with the following properties of a protein: its cellular location, biological function, folding type and disulfide bonding. The composition of an extra-cellular protein, which is characterized by harboring a signal peptide at the N-terminus and subsequently processed, was different from that of intracellular proteins.

These observations showed that intracellular and extra-cellular proteins have different amino acid compositions and paved the way for discerning their location solely from compositional data (Nakashima & Nishikawa 1992). Taking these cues a few steps further, membrane proteins were examined to identify differences in amino acid composition between cytoplasmic and extra-cellular sides. Peptide chains could be correctly assigned as either cytoplasmic or extra-cellular solely from sequence composition. For single spanning membrane proteins the predictive accuracy was 90%, whereas for multi-spanning proteins it was 85% (Nakashima & Nishikawa 1992).

Subsequent detailed examination for prominent patterns revealed that the intracellular proteins are relatively rich in aliphatic as well as charged residues. A transporting polypeptide having successive hydrophobic residues is trapped by the membrane to become a transmembrane protein, and that charged residues (particularly lysine and arginine) tend to stop the transport of a polypeptide (Von Heijne 1992). In this connection, the smaller amount of hydrophobic and charged residues observed in extra-cellular proteins seem to be more ideally suited for transport across the membrane. (Nakashima & Nishikawa 1994).

Subsequently Cedano et al. (1997) developed an algorithm *ProtLock* to predict the cellular location of a protein. Predictions by *ProtLock* are based on computing amino acid frequencies of a protein and measuring its distance (statistical measure Mahalanobis D^2 distance) between the query protein and a reference set culled from SWISS-PROT database. Query proteins are assigned to the class to which the resulting distance is minimum. Prediction exercises using *ProtLock* have relatively high efficiency (in the range 70%-80%). More recently Yu et al. (2004) have used support vector machines based on n-peptide compositions for predicting sub-cellular localizations of proteins with an efficiency of 89%. Support Vector Machines (SVM, Vapnik 1995) is now being increasingly applied in bioinformatics

due to its reported ability to efficiently learn from a reference set and assign a query protein to any of the classes. Although these methods have relatively high efficiency, it has also come to light that, used in this form, compositional analysis is unable to yield 100% accuracy. The reasons presumably reside in the possibility of other players such as the specific order of certain amino acids, accumulation of conservative substitutions and as yet unknown factors that may guide proteins to their correct sub-cellular destinations. Alternatively compositional analysis needs exploring other formats of symbol mapping techniques for example, the proportion of amino acids classified according to hydrophobic scales and so on. Nonetheless, it is now well established that amino acid composition of a protein is a major molecular determinant of sub-cellular location of proteins.

Stability

The factors affecting protein stability are pH, ligand binding, disulphide bonds, and the role of individual amino acid residues. Guruprasad and co-workers (1990) made an early observation on the correlation of amino acid composition with *in vivo* protein stability. Subsequently Reddy re-examined this aspect on a later dataset (1996). It was observed that the primary determinants of the stability of a protein reside in its primary structure and is an intrinsic property of the protein. The presence of certain dipeptides and their frequency of occurrence render a protein stable or susceptible to degradation. Reddy classified 102 dipeptides that were found to play significant role in determining the intracellular protein stability into stable, destabilizing and normal classes. The stabilizing dipeptides consisted more of hydrophobic combinations whereas the destabilizing dipeptides were found to be more of hydrophilic combinations. The combination of glycine, threonine and valine were predominantly of stable class whereas combinations of aspartate, cysteine, methionine, arginine and serine were of destable class. In summary, the occurrence of certain dipeptides was significantly different in unstable proteins compared to stable polypeptides. However, cross-linking between cysteine groups might override the impact of dipeptides. Thus a careful consideration of multiple factors is suggested in addition to determining the dipeptide composition to assess the overall *in vivo* stability of a given protein.

Structure

Most proteins have roughly spherical shapes and are generally referred to as globular. They are soluble and can be crystallized and studied experimentally. On the other hand, fibrous proteins play structural roles and have regular, extended structures. The large sizes of fibrous proteins and their general insolubility make them more difficult to be characterized experimentally. However, advances have been made in this realm also by choosing shortened version of these proteins for experimental analysis. One of the first proteins to be structurally characterized was collagen. The triple helical collagen has three polypeptide chains with a large number of repeat sequences of the type Gly-X-Y where X is often proline and Y is often hydroxyproline. The triple helical arrangement imposes the critical requirement that every third residue must be a glycine. This sequence requirement is a hallmark of triple helix collagen-like domains. Since this discovery, parallel clusters of short triple helices were also observed in other proteins such as blood complement component C1q, Acetylcholinesterase, Fibronectin, Osteonectin, and sugar-binding collectins (Smith 1986, Håkansson & Reid 2000).

Several structural proteins involved in maintaining cell shape, organizing cytoplasm and serving locomotion have coiled coil shapes, which are two or three α-helices wound around each other to form a left-handed super-helix. Examination of the sequences of these proteins revealed that these segments consist of regular patterns in their amino acid sequences. These sequences are repetitive with a period of seven residues and hence the term 'heptad repeat' is used to signify them. During the formation of a coiled-coil structure from α-helices the side chains with similar physico-chemical properties pack against each other.

Heptad repeats occur in many proteins with diverse functions. Examples are motor protein myosin, fibrinogen, collectins (cell-surface recognition proteins), spectrin and dystrophin (Adamson et al. 1993). In some cases, a special class of heptad repeats consisting of leucine zipper occurs, for example, the dimerization of transcription factor GCN4 is accomplished by the formation of α-helical coiled-coil with leucine zipper (Alber 1992). Similar mechanism underlies the cause of oligomerization of the leucine zipper heptad repeat of gp41 transmembrane protein of HIV-1 (Human Immunodeficiency Virus type 1; Bernstein et al. 1995).

Silk fibroins on the other hand belong to a different group of structural proteins. The protein is composed of β-sheets (parallel or anti-parallel arrangement remains to be resolved) consisting of Gly-X (X = A, S or Y) type repeats covering a major portion of the sequence (Zhou et al. 2001). Amyloid fibrils also have β structure (Sipe & Cohen 2000).

The regularity in sequence patterns along with the imposition of a structural constraint that only certain amino acids are accepted in the reiterated pattern dictates a distinct compositional requirement characteristic of several structural proteins. This aspect has become to be known as 'compositional bias'. Compositionally biased sequences are also known as simple sequences or sequences of low complexity.

Proteins with Compositional Bias

Simple sequences: a Generalization of Repeats

The advent of genomics era during mid-90s fuelled this further. Concomitantly, the analysis of protein sequences although progressing slowly, revealed a variety of simple sequence patterns in proteins. Simple sequences can be broadly classified into three types: (i) cryptically simple, (ii) regular reiterated pattern, (iii) regular runs (figure 1). The examples of structural proteins described in the previous section have regular reiterated patterns. In natural sequences, simple and complex segments occur juxtaposed revealing an underlying mosaic structure. Soding and Lupas (2003) have proposed that modern proteins evolved by fusion and recombination from a more ancient peptide world. It is probable that some peptides were of simple sequence type whereas others were of complex sequence type so that final fused functional polypeptide is a mosaic of simple and complex sequence types. Because complex sequences carry more information compared to simple sequences, the latter were not the focus of general sequence analysis and were relegated to low priority for a long time.

Recently, a great surge in interest in analyzing the distribution of simple sequences in proteins has arisen principally because they were observed to be associated with human neuro-degenerative disorders (Bhandari & Brahmachari 1995, Karlin et al. 2002). For example, the huntingtin protein associated with Huntington disease contains long stretches of glutamine Q_{23}, Proline P_{11}, P_{10} and glutamic acid E_5, E_6. Atropin-1 associated with dentatorubral pallidoluysian atrophy, DRPLA contains Q_{20}, S_7, S_{10}, P_6, H_5; the androgen receptor protein (associated with Kennedy's disease) contains Q_{26}, Q_6, Q_5, P_8, A_5, G_{24}; and the brain voltage-dependent calcium channel protein CCAA (spino-cerebellar ataxia 6) contains H_{10}, Q_{11}. PolyAlanine expansions also result in a variety of diseased conditions in humans (Cocquet et al. 2003). A more general survey of polyAlanine containing proteins showed that these proteins also contained runs of other amino acids such as Histidine, Serine, Glutamine and Proline (Veitia 2004).

Proline rich sequences are known for its structural role mediating protein-protein interactions (Oneyama et al. 2003). Sequences enriched in charged residues are associated with DNA and RNA processing, chromatin structures, ion-binding and protein-protein interactions (Davis et al. 1998). Tracts of poly-glutamine may mediate protein-protein interactions and arginine rich regions have been found to be involved in protein-RNA interaction (Karlin et al. 2002, Wilkinson et al. 2000, Poisson et al. 1995). The association of simple sequences with several human genetic diseases and their observed structural roles in specific context of biological processes such as transcription regulation and protein-protein interaction has stimulated rigorous analysis of simple sequences. Analysis of simple sequences is important to reveal their evolutionary history, molecular mechanisms of generating variants and selective over representation in selected functional classes of proteins. We were stimulated to analyze complete proteome sequences to answer the following fundamental questions:

(a) What is the nature of simple sequences and which amino acids are represented in them?

(b) What is the distribution of simple sequences in functional classes?

Simple Proteins in Complete Proteomes

Traditionally, complex sequences had received attention for identifying biochemical functional motifs. A complex protein sequence for example, can carry instruction for folding and packing, for correct stereo-chemical positioning of motifs, for organization into domains, for solubility and for determining overall shape all in one sequence. Because simplicity is antipode of complexity, analysis of simple sequences was carried out from the opposite angle. Algorithms were first developed to assess the

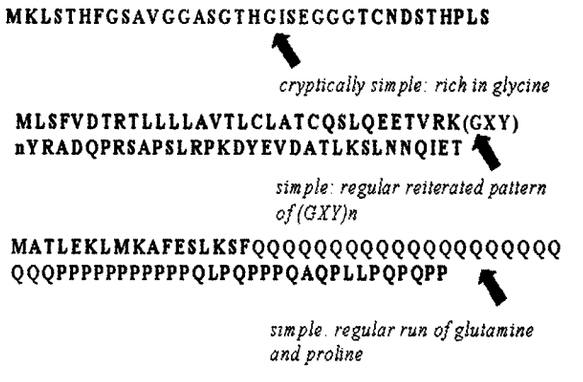

Figure 1. Trinity of Simple sequences in proteins. Based on several surveys

complexity of a given sequence. Sequences that did not cross a preset threshold value were classified as 'low complexity' that formed the seed for analysis of simple sequences. Only more recently, algorithms have been developed that directly focus on identifying simple sequences and analyzing them.

Rigorous formalisms of sequence complexity with correlation of function require the synthesis of quantitative expressions estimating sequence complexity and function or structure of a molecule. Representation of biological function in mathematical terms, although attempted, is not easy. In contrast, analysis of structure can be carried out in mathematical terms because of the possibility of applying geometrical principles to quantify structural characteristics of a molecule.

Early on, Wootton tackled the issue of sequence complexity from the perspective of information theory (Wootton 1994). The algorithm 'SEG' born out of this effort is still in use today as default filter in BLAST searches to mask simple sequences to reveal true orthologous relationships. Application of SEG revealed that compact globular domains (spherical type) of proteins potentially forming highly ordered crystal structures have high sequence complexity of amino acid composition. Conversely, non-globular domains (extended type) had significantly lower compositional complexity (Wootton 1994). In other words, simple sequences were composed of reiterative motifs. These observations were consistent with observed data on several structural proteins in that, repeated sequences adopt non-globular (rod-like or sheet-like) structures.

Popov and Trifonov (1996) developed a measure for estimation of sequence complexity from a linguistic perspective. The proposed measure computes the ratio of number of unique vocabulary of words in a given segment of the sequence to the maximum possible vocabulary for the same composition of amino acids or nucleotides. Application of this linguistic measure to nucleotide sequences was helpful in identifying the correlation of nucleosome forming sequences with differing levels of sequence complexity (Bolshoy et al. 1997). The latter is computationally simple compared to SEG.

At the time when we had initiated analysis of simple sequences about 4 years ago, these were the two algorithms that were either widely used or applied to extract useful biological information from molecular sequences. Both algorithms were not suitable for comparative proteomics analysis. SEG is accompanied with the limitation that a given run is subject to a set of 3 parametric values and therefore practical considerations demand that comparative analysis be restricted to only one set of parameters. The measure proposed by Bolshoy et al. (1997) assesses the complexity of a sequence with respect to the maximal possible vocabulary size computed assuming an infinite occurrence of the monomer units of same composition. Although both measures have been used and have provided some useful results, inherently they do not consider the practical and finite aspects of biological sequences.

We therefore developed a novel algorithm *ScanCom* to assess sequence complexity of proteins based on a novel dimer word count approach (Nandi et al. 2003a) that overcomes the limitations associated with previous algorithms. *ScanCom* operates with one parameter, namely, the size of the sliding window to scan the sequences. Second, our algorithm is simple and easy to comprehend. The dimer (dipeptide) word count principle has been already established as a molecular determinant of several features of proteins such as stability and spatial structure. Third, *ScanCom* takes into account the finite aspects of the sequence. Fourth, we applied the measure to test cases and determined the demarcation line between complex and simple sequences attached with biological significance from a structural angle. We determined an optimal value for the sliding window (45 amino acids) and a demarcation value (F_c (proportion of reiterated dipeptides in a protein) $\geq$ 15) for partitioning proteins into either high complexity (complex) or low complexity (simple) categories using test cases of high-resolution crystal structures.

Application of *ScanCom* to 38 complete proteome sequences of bacteria available at the time partitioned proteins into complex and simple categories. A striking feature readily emerging from this analysis revealed that in most bacteria the fraction of simple sequence containing proteins comprised of at most 7% and varied from one species to another. Five organisms had high proportion of simple proteins: the archeaon *Aeropyrum pernix*, and the eubacteria *Deinococcus radiodurans* (radiation resistant), *Caulobacter crescentus* (a plant pathogen), *Mycobacterium tuberculosis* (a human pathogen), and *Pseudomonas aeruginosa* (an opportunistic pathogen of humans, causing a range of pathology).

Detailed examination of simple proteins from these organisms unfolded the types of reiterated sequences contained in them. There were repeats with clusters of charged amino acids, regular patterns and regular runs of amino acids, and several cryptically simple motifs (Nandi et al. 2003a). The amino acids Leucine, Alanine, Glycine and Proline were top ranking in the repeats. Alanine, Glycine and Proline were classified as 'early evolved' by Trifonov on the basis

of forty criteria (Trifonov 2000). Leucine was classified as amino acid with simple structure and was observed in Miller's mixture emulating pre-biotic soup of Earth.

These observations led us to propose that repeated sequences in bacterial proteins primarily evolved from expandability of earliest codons. Therefore, the expansion of triplet repeats *per se* evolved very early in evolution and is not peculiar to mammalian or primate lineages. However, the type of codons undergoing expansion in humans and implicated in human genetic diseases code for amino acids that evolved later such as glutamine. Expansions of codons coding for proline and other 'early evolved' amino acids in the human genome may represent evolutionary relics.

Ever since the subject of repeats became the focus of frontline investigations it was becoming apparent that simple proteins had restricted representation in the wide array of functional activities that proteins carry out in a cell. Smooth regular runs of amino acids were frequent in transcription factors and other proteins that participate in protein-protein interactions for accomplishing the desired function. However, a systematic investigation of the distribution of simple proteins in functional classes has not been undertaken so far. In an effort to open this 'Pandora's box', we designed a novel classification scheme (Nandi et al. 2003b).

The prime mover for designing this scheme was the preliminary observation of very low number of simple proteins in each functional class according to the existing scheme proposed by Riley (1993). The classification we adopted was agglomerative type, which consisted of lumping the individual functional classes into two super-classes: CELLULAR PROCESSES (CP) and TRANSPORT and MEMBRANE ASSOCIATED (TM) based on annotation provided in the sequence files. The remaining proteins were collectively binned into a separate class termed CHARACTERISTIC. Several proteins of this super-class have functional roles that confer unique biological property on the species under study.

In an exercise of comparative analysis of simple proteins from *Escherichia coli* K12, *E. coli* O157, *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 and *M. leprae*, a striking observation was made from macroscopic angle (Nandi et al. 2003c). The proportion of TRANSPORT and MEMBRANE ASSOCIATED simple proteins in Mycobacteria was lower compared to Escherichia mirroring a previous analysis (Tekaia et al. 1999). More importantly, we observed that although *M. leprae* has a drastically reduced proteome size compared to *M. tuberculosis*, the former does not compromise on the number of simple transport and membrane proteins thereby indicating that we had struck at a 'minimal set' of these proteins essential for growth and physiology of Mycobacteria.

Further, we could readily point out that the pathogenic strain of *E. coli* O157 differed from the non-pathogenic strain *E. coli* K12 in having acquired simple protein directly functioning in pathogenic processes (Nandi et al. 2003c). These differences between species and between strains of the same species were particularly notable in the CHARACTERISTIC super-class. The same scheme of comparative proteomics using simple proteins produced interesting results in a comparative analysis of enteric bacterial pathogens: *E. coli* O157 (enteropathogenic), *Vibrio cholerae* (causes diarrhoea), *Helicobacter pylori* (causes gastric ulcer) and *Campylobacter jejuni* (food borne pathogen). We could clearly identify that greater similarity exists among simple proteins of CELLULAR PROCESSES class between *E. coli* and *V. cholerae* and between *H. pylori* and *C. jejuni* than in other comparisons.

These observations parallel taxonomic classification because *E. coli* and *V. cholerae* are classified under gamma sub-division of proteobacteria whereas *H. pylori* and *C. jejuni* are classified under epsilon sub-division of proteobacteria. Simple proteins of other classes displayed more variability and no pattern was discernible.

These results uncovered the differences in rates of evolution of simple proteins from different compartments of bacterial proteomes. This aspect is consistent with the notion that proteins of the CELLULAR PROCESSES class are more ancient, function intracellularly, and have evolved by vertical descent. In contrast, proteins of TRANSPORT and MEMBRANE ASSOCIATED class function at the interface between a cell and its niche and may have evolved variations through natural selection to suit the specific requirements (Fraser 2000). Proteins of CHARACTERISTIC class in general have functional roles in conferring unique biological characteristics and therefore are likely to be most divergent in inter-species and perhaps even in inter-strain comparisons.

Emerging Themes and Applications

New Facets of Simple Sequences

As mentioned earlier, contemporaneously to our efforts, a few other groups developed new algorithms to examine simple sequences somewhat directly. Promponas et al (2000) have recently developed an algorithm CAST (Complexity Analysis of Sequence Tracts) to specifically detect runs of single residue

types. CAST can also be used to identify and generate a dataset of compositionally biased regions forming a crude seed for further analysis. Romero et al. (1999) developed two measures to estimate sequence complexity, namely, sequence entropy and alphabet size with a focus on identifying lower bound for sequence complexity for the folding of globular proteins into domains with biological function. It was pointed out that, there is a local complexity requirement for binding pocket formation, a global complexity requirement for rigid packing and for surface characteristics leading to solubility. Further, these authors observed that disordered regions in several proteins were of 'simple' sequence type.

Mar Alba and coworkers (2002) oriented the SIMPLE algorithm to search for cryptically simple sequences in proteins. Analysis of yeast proteins using this algorithm identified over-represented and not so abundant amino acids in cryptic simple sequences. These were classified into functional classes. For example, cryptic simple repeats in permeases had isoleucine in high representation whereas the repeats in protein kinases had serine and glutamine in abundance. These measures had an in-built deficiency in that the crystal structure data used to assess their measures lacked a rigorous mathematical framework for quantitative comparison in contrast to our algorithm. This brings into question the full biological significance of these findings. Nonetheless, these efforts attest the resurgence of interest in analyzing and characterizing simple sequences in proteins.

Elisabetta Pizzi and Clara Frontali (2001) carried out a detailed analysis on the protein sequences of two chromosomes (2 and 3) of *Plasmodium falciparum*. These authors have used the algorithm developed by Wootton and Federhen (1996) to segregate low complexity segments from high complexity regions. *P. falciparum* causes the dreaded disease malaria, predominantly in developing countries and therefore analysis of its sequences could reveal insights into the molecular features of its proteins. A large number of proteins of *P. falciparum* are unusual in that they contain long segments of runs of asparagine residues. An important observation emerging from this analysis suggested that the driving force of compositional features of simple sequence segments is the high A+T content of the codon vocabulary used by *P. falciparum*. Both substitutions and expansion/contraction sequence polymorphisms were found to occur in the Plasmodia lineage. The simple sequence segments were generally highly hydrophilic.

Marcotte and coworkers (1998) recently documented a census of repeated sequences in representatives from all the three primary kingdoms of life: Archaea, Eubacteria and Eukaryota. They report that repetitive sequences are more common in eukaryotic proteins than in prokaryotic proteins. Further, their observations suggest that proteins with repetitive sequences evolve faster than those of high sequence complexity. These authors also find that similar repeat forming mechanisms are operating in the different kingdoms of life although the variety of repeats in eukaryotes is higher than in prokaryotes. Our independent observations are consistent with these general postulations.

For several decades the axiom *Amino acid sequence → 3 Dimensional structure → Function* has dominated the analytical framework of biochemists and molecular biologists. Recently, it has been recognized that numerous proteins lack intrinsic globular structure or they contain long disordered segments under physiological conditions (Wright & Dyson 1999, Dunker & Obradovic 2001). More interestingly, this is their normal, functional state. Such proteins were found to be frequently involved in regulatory functions in the cell and, a structure is adopted when the protein binds to its target molecule. It was observed that many disordered segments of proteins indeed have a simple sequence structure and these intrinsically unstructured proteins evolve by repeat expansion (Tompa 2003).

The lack of intrinsic structure has been hypothesized to confer the advantage of flexibility to bind to several targets instead of having several individual proteins to accomplish the same functional event. In this sense it appears that in these regulatory proteins, a 'clay-mould' (adaptability) model was selected in favor compared with the 'lock and key' (fixed) model. In the 'clay-mould' model, a simple sequence (disordered) segment is analogous to clay and moulds are structured (ordered) regions in the targets to which the disordered region of the partner molecule will bind. Adoption of the 'lock and key' model by a cell enables a one-to-one binding to turn on a function whereas the 'clay-mould' model provides a larger space by effecting a one-to-many binding to turn on or aid several functions if required. The latter could be advantageous in regulatory switching wherein a simultaneous trigger can aid rapid transmittance of regulatory signals.

In this context, it is not surprising that many transcription factors have simple sequence segments presumably disordered but with potential to bind DNA and participate in protein-protein interactions. However, the fact that these polypeptide segments are composed of covalently linked monomers raises a suspicion: can these sequences ever adopt regular

conformations? Sharma et al. (1999) illuminated this point by demonstrating that polyglutamine sequences, which are the most commonly observed simple sequences in transcription factors, adopt a β sheet conformation. Thus it is possible that simple sequences may oscillate between a fully disordered and an entirely structured conformation. However, a lot of work remains to be done before we begin to comprehend the precise determinants of a regular conformation *vis-à-vis* a disordered flexible state.

The adoption of the 'clay-mould' model is accompanied by a caveat: the distribution of simple sequences of certain types must be restricted to that population of proteins which requires flexible portions and brings about a given event through a labyrinth of interactions. Indeed, surveys of regular runs of amino acids in human proteins are consistent with this hypothesis (Bhandari & Brahmachari 1995, Karlin et al. 2002). In the context of human genetic diseases it appears that, acquiring flexible regions by accumulating simple sequences was accompanied with the danger of potential expandability of the codons coding for these sequences through slippage during replication or unequal crossing-over mechanisms. A diseased condition is precipitated when expansions cross a certain threshold.

New Facets of Compositional Analysis

Recently several authors have used compositional analysis to unravel interesting biological features. For example, Schneider (1999) developed an artificial neural network based approach using amino acid composition to predict secreted proteins in bacterial proteomes by taking advantage of the observations made by Nakashima and Nishikawa (1994) and Cedano et al. (1997). Development of this algorithm enabled short-listing potentially secreted proteins. In the context of pathogens, they could be further characterized for identifying new drug targets.

Another implementation of compositional analysis was in the development of the algorithm PATS (Predict Apicoplast-Targeted Sequences) by Zuegge et al. (2001). This algorithm allows the prediction of apicoplast targeted proteins in *P. falciparum*. Apicoplast is an organelle of prokaryotic origin and therefore the proteins of this organelle offer attractive prospects for developing new drug targets because chloroquine resistant strains have now become widespread.

Hobohm and Sander (1995) developed a property-based approach to search protein databases. The algorithm PropSearch uses 144 compositional properties and is useful where the traditional sequence alignment algorithms such as BLAST fail to produce unequivocal results. Recently we have developed an algorithm SPAAN (Software for Prediction of Adhesins and Adhesin-like proteins using Neural networks) to predict adhesins and adhesin-like proteins in bacterial pathogens (Sachdeva et al. 2005). The algorithm uses 105 compositional properties combined with the power of neural networks. Because the first step in a host-pathogen interaction is mediated by adhesins, the identification of novel adhesins in pathogenic microbes can accelerate the formulation of new vaccines to be tested in suitable animal models.

Future Prospects

Compositional analysis of sequences, an age-old platform has now arrived at the forefront of analysis of proteins primarily because of the limitations of the existing sequence alignment based algorithms for prediction of function. When sequence alignment based approaches were developed, they appeared very powerful and indeed revolutionized experimental molecular biology. Today, the genomic revolution has thrown up a new challenge forcing investigators to combine old and new approaches together. The various approaches attempted by combining basic compositional computations with neural networks or support vector machines have opened a new field. Other numerous approaches including different combinations of symbol mapping techniques and statistical formulations need to be tried to push the power of predictions further. This field is now likely to develop further and new applications will be invented. The resurgence of interest in simple sequences has spread worldwide. More interestingly newer themes are evolving around simple sequences: the trinity of protein structure (ordered structure, molten globule and random coil), disordered segments and protein function, comparative proteomics analysis to reveal biological features of pathogens and illumination of the evolution of mechanistic aspects of codon expansion in human genome. As I see it, Bioinformatics approach is now becoming increasingly Bohemian: a blend of old and new approaches will be taken not necessarily in any particular order.

Acknowledgements

I would like to specially thank Prof. Samir K. Brahmachari, whose obsession with structural and functional characterization of reiterated sequences had a positive inductive effect on me. I thank Ms. Tannistha Nandi, Prof. C. Ramakrishnan, Emeritus Professor IGIB, CSIR for their excellent contributions, Mr. Mandapati Kiran Kumar and Ms. Anannya Bandyopadhyay for help with preparing the manuscript, and The Council of Scientific and Industrial Research and Department of Biotechnology, Govt. of India for funding support.

References

Adamson J G, Zhou N E, Hodge R S 1993 Structure, function and application of the coiled-coil protein folding motif; *Curr. Opin. Biotechnol.* **4** 428-437

Alber T 1992 Structure of the leucine zipper; *Curr. Opin. Genet. Dev.* **2** 205-210

Altschul S F, Gish W, Miller W, Myers E W, Lipman D J 1990 Basic local alignment search tool; *J. Mol. Biol.* **215** 403-410.

Bernstein H B, Tucker S P, Ka, S R, McPherson S A, McPherson D T, Dubay J W, Lebowitz J, Compans R W, Hunter E 1995 Oligomerization of the hydrophobic heptad repeat of gp41; *J. Virol.* **69** 2745–2750.

Bhandari R, Brahmachari S K 1995 Analysis of CAG/CTG triplet repeats in the human genome: implication in transcription factor gene regulation; *J. Biosc.* **20** 613-627.

Boeckmann B, Bairoch A, Apweiler R, Blatter M.-C, Estreicher A, Gasteiger E, Martin M J, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M 2003

The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003; *Nucleic Acids Res.* **31** 365-370.

Bolshoy A , Shapiro K , Trifonov E N, Loshikhes I 1997 Enhancement of the nucleosomal pattern in sequences of lower complexity; *Nucl. Acids Res.* **25** 3248-3254.

Cedano J, Aloy P, Perez-Pons, J A, Querol, E 1997 Relation between amino acid composition and cellular location of proteins; *J. Mol. Biol.* **266** 594-600.

Cocquet J, De Baere E, Caburet S, Veitia R A 2003 Compositional Biases and Polyalanine Runs in Humans; *Genetics* **165** 1613–1617.

Davis S J , Davies E A , Tucknott M G , Jones E Y, van der Merwe P A 1998 The role of charged residues mediating low affinity protein-protein recognition at the cell surface by CD2; *Proc. Natl. Acad. Sci. USA* **95** 5490-5494.

Dunker A K, Obradovic Z 2001 The protein trinity-linking function and disorder. *Nat. Biotechnology* **19** 805-806.

Fraser C M, Eisen J, Fleischmann, R D, Ketchum K A, Peterson S 2000 Comparative genomics and understanding of Microbial Biology; *Emerg. Infect. Dis.* **6** 505-572.

Guruprasad K, Reddy B V, Pandit M W 1990 Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence; *Protein Eng.* **4** 155-161.

Håkansson K, Reid K B M 2000 Collectin structure: A review; *Prot. Sci.* **9** 1607–1617.

Hobohm U, Sander C, 1995 A sequence property approach to searching protein databases; *J. Mol. Biol.* **251** 390-399.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles, A J 2002 Amino acid runs in eukaryotic proteomes and disease associations; *Proc. Natl. Acad. Sci. USA* **99** 333-338.

Mar Alba M, Laskowski R A, Hancock J M 2002 Detecting cryptically simple protein sequences using the SIMPLE algorithm; *Bioinformatics* **18** 672-678.

Nandi T, Dash D, Ghai R, B-Rao C, Kannan K, Brahmachari S K, Ramakrishnan C, Ramachandran S 2003a A novel complexity measure for comparative analysis of protein sequences from complete genomes; *J. Biomol. Str. Dyn.* **20** 657-667.

------, Kannan K, Ramachandran S 2003b The low complexity proteins from enteric pathogenic bacteria: taxonomic parallels embedded in diversity; *In Silico Biology* **3** 277-285.

Nandi T, Kannan K, Ramachandran S 2003c Species and strain-specific patterns of low-complexity proteins in Escherichia and Mycobacteria; *Curr. Sci.* **85** 185-187.

Oneyama C, Agatsuma T, Kanda Y, Nakano H, Sharma S V, Nakano S, Narazaki F, Tatsuta K 2003 Synthetic inhibitors of proline-rich ligand-mediated protein-protein interaction: potent analogs of UCS15A; *Chem. Biol.* **10** 443-451.

Nakashima H, Nishikawa K 1992 The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins; *FEBS Lett.* **303** 141-146.

------, Nishikawa K 1994 Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies; *J. Mol. Biol.* **238** 54-61.

Perutz M 1992 *Protein Structure New approaches to disease and therapy* Freeman and Company, New York, pp 257-260.

Popov O, Segal D M, Trifonov E N 1996 Linguistic complexity of protein sequences as compared to texts of human languages; *Biosystems.* **38** 65-74.

Poisson F, Roingeard P, Gourdeau A 1995 Direct investigation of protein RNA binding domains using digoxigenin-labelled RNAs and synthetic peptides: application to the hepatitis delta antigen; *J. Virol. Methods* **55** 381-389.

Promponas V J, Enright A J, Tsoka S, Kreil D P, Leroy C, Hamodrakas S, Sander C, Ouzounis C A 2000 CAST: an iterative algorithm for the complexity analysis of sequence tracts; *Bioinformatics* **16** 915-922.

Ramachandran G N and Kartha G 1955 Structure of Collagen; *Nature* **176** 593-595.

Reddy B V 1996 Structural distribution of dipeptides that are identified to be determinants of intracellular protein stability; *J. Biomol. Struct. Dyn.* **14** 201-210.

Riley M 1993 Functions of the gene products of *Escherichia coli*; *Micobiol. Rev.* **57** 862-952.

Romero P, Obradovic Z, Dunker A K 1999 Folding minimal sequences: the lower bound for sequence complexity of globular proteins; *FEBS Lett.* **462** 363-367.

Sachdeva G, Kumar K, Jain P, Ramachandran S 2005 Software for Prediction of Adhesins and Adhesin-like proteins using Neural networks.

------ 1999 How many potentially secreted proteins are contained in a bacterial genome?; *Gene* **237** 113-121.

Sharma D, Sharma S, Pasha S, Brahmachari S K 1999 Peptide models for inherited neurodegenerative disorders: conformation and aggregation properties of long polyglutamine peptides with and without interruptions; *FEBS Lett.* **456** 181-185.

Sipe J D, Cohen A S 2000 Review: history of the amyloid fibril; *J Struct .Biol.* **130** 88-98.

Smith R 1986 The molecular genetics of collagen disorders. *Clinical Science* **71** 129-135.

Soding J and Lupas A N 2003 More than the sum of their parts: on the evolution of proteins from peptides; *BioEssays* **25** 837-846.

Tekaia F, Gordon S V, Garnier T, Brosch R, Barrell B G, Cole S T 1999 Analysis of the proteome of *Mycobacterium tuberculosis* in silico; *Tuber Lung Dis* **79** 329-342.

Tompa P 2003 Intrinsically unstructured proteins evolve by repeat expansion; *BioEssays* **25** 847-855.

Trifonov E N 2000 Consensus temporal order of amino acids and evolution of the triplet code; *Gene* **261** 139-151.

Vapnik V 1995 The nature of statistical learning theory; Springer, New York.

Veitia RA 2004 Amino acids runs and genomic compositional biases in vertebrates; *Genomics* **83** 502-507.

Von Heijne G 1992 Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule; *J. Mol. Biol.* **225** 487-494.

Wilkinson T A, Botuyan M V, Kaplan B E, Rossi J J and Chen Y 2000 Arginine side-chain dynamics in the HIV-1 rev-RRE complex; *J. Mol. Biol.* **303** 515-529.

Wootton J C 1994 Non globular domains in protein sequences: automated segmentation using complexity measures; *Comput. Chem.* **18** 269-285.

Wootton J C and Federhen S 1996 Analysis of compositionally biased regions in sequence databases; *Methods Enzymol.* **266** 554-571

Wright P E and Dyson H J 1999 Intrinsically unstructured proteins: re-assessing the protein structure –function paradigm; *J. Mol. Biol.* **293** 321-331.

Zhou C Z, Confalonieri F, Jacquet M, Perasso R, Li Z G and Janin J 2001 Silk fibroin: structural implications of a remarkable amino acid sequence; *Proteins* **44** 119-22.

Zuegge J, Ralph S, Schmuker M, McFadden G I and Schneider G 2001 Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins; *Gene* **280** 19-26.