45. van der Bliek, A. M. & Meyerowitz, E. M. Dynamin-like protein encoded by the *Drosophila shibire* gene associated with vesicular traffic. *Nature* **351**, 411–414 (1991).
46. Chen, M. S. *et al.* Multiple forms of dynamin are encoded by *shibire*, a *Drosophila* gene involved in endocytosis. *Nature* **351**, 583–586 (1991).
47. Praefcke, G. J. & McMahon, H. T. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nature Rev. Mol. Cell Biol.* **5**, 133–147 (2004).
48. Robinson, M. S. Adaptable adaptors for coated vesicles. *Trends Cell Biol.* **14**, 167–174 (2004).
49. Sorkin, A. Cargo recognition during clathrin-mediated endocytosis: a team effort. *Curr. Opin. Cell Biol.* **16**, 392–399 (2004).
50. Ehrlich, M. *et al.* Endocytosis by random initiation and stabilization of clathrin-coated pits. *Cell* **118**, 591–605 (2004).
51. Wu, X. *et al.* Clathrin exchange during clathrin-mediated endocytosis. *J. Cell Biol.* **155**, 291–300 (2001).
52. Gaidarov, I., Santini, F., Warren, R. A. & Keen, J. H. Spatial control of coated-pit dynamics in living cells. *Nature Cell Biol.* **1**, 1–7 (1999).
53. Rappoport, J. Z., Taha, B. W. & Simon, S. M. Movement of plasma-membrane-associated clathrin spots along the microtubule cytoskeleton. *Traffic* **4**, 460–467 (2003).
54. Merrifield, C. J., Feldman, M. E., Wan, L. & Almers, W. Imaging actin and dynamin recruitment during invagination of single clathrin-coated pits. *Nature Cell Biol.* **4**, 691–698 (2002).
55. Marsh, M. & McMahon, H. T. The structural era of endocytosis. *Science* **285**, 215–220 (1999).
56. Fotin, A. *et al.* Molecular model for a complete clathrin lattice from electron cryomicroscopy. *Nature* **432**, 573–579 (2004).
57. Barbieri, M. A., Ramkumar, T. P., Fernadez-Pol, S., Chen, P. I. & Stahl, P. D. Receptor tyrosine kinase signaling and trafficking paradigms revisited. *Curr. Top. Microbiol. Immunol.* **286**, 1–20 (2004).
58. Pelkmans, L. *et al.* Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature* **436**, 78–86 (2005).
59. Roth, M. G. Phosphoinositides in constitutive membrane traffic. *Physiol. Rev.* **84**, 699–730 (2004).
60. Stahl, P. D. & Barbieri, M. A. Multivesicular bodies and multivesicular endosomes: the 'ins and outs' of endosomal traffic. *Sci. STKE* **2002**, PE32 (2002).
61. Marsh, M. & Thali, M. HIV's great escape. *Nature Med.* **9**, 1262–1263 (2003).
62. Russell, D. W. *et al.* cDNA cloning of the bovine low density lipoprotein receptor: feedback regulation of a receptor mRNA. *Proc. Natl Acad. Sci. USA* **80**, 7501–7505 (1983).
63. Xie, X. S., Stone, D. K. & Racker, E. Activation and partial purification of the ATPase of clathrin-coated vesicles and reconstitution of the proton pump. *J. Biol. Chem.* **259**, 11676–11678 (1984).
64. Forgac, M. & Cantley, L. Characterization of the ATP-dependent proton pump of clathrin-coated vesicles. *J. Biol. Chem.* **259**, 8101–8105 (1984).

Competing interests statement
The author declares no competing financial interests.

DATABASES
The following terms in this article are linked online to:
OMIM:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
hypercholesterolaemia
Swiss-Prot: http://www.expasy.org/sprot
AP50 | (LDL)-receptor | Shibire

FURTHER INFORMATION
Michael Roth's laboratory: http://www.utsouthwestern.edu/findfac/research/0,2357,16262,00.html
Access to this interactive links box is free online.

OPINION

Protein linguistics — a grammar for modular protein assembly?

Mario Gimona

Abstract | The correspondence between biology and linguistics at the level of sequence and lexical inventories, and of structure and syntax, has fuelled attempts to describe genome structure by the rules of formal linguistics. But how can we define protein linguistic rules? And how could compositional semantics improve our understanding of protein organization and functional plasticity?

Post-genomic activities have flooded databases and the scientific literature with new protein entries. These identifications add to the lexical complexity of the proteome, but what about functional (semantic; see glossary box for definitions of linguistic expressions) annotation? Even SMART (see Online links box), the schematic domain-representation system, falls short of supplying precise functional descriptions. This is mostly because the literature on proteins and protein modules in the database is highly ambiguous with respect to their reported functional properties.

And for the most part, the lack of unambiguous functional annotation causes considerable confusion and misapprehension when attempting to extrapolate the properties of any given protein on the basis of those functions that are portrayed for domain constituents that were described in other (homologous) molecules. Often, the results of such 'homology strategies' resemble those of early endeavours towards reconstructing the history of languages by comparing different lexica (see REF. 1 for an example). Notably, this problem has been solved for human languages by looking at the underlying principles of grammar, and even distant languages are similar with respect to the grammatical rules that apply to them[2].

Several attempts have been made over the past three decades to combine linguistic theory with biology[3–6], which led researchers to propose terms like "cell language"[7,8], to establish parameters for general "molecular linguistics"[9], or to suggest "folding as grammar"[10]. Over 20 years ago, Brendel and Busse[11,12] used formal linguistic concepts to define a basic set of grammatical rules for genes, based on the idea that mutating a piece of genetic information was similar to modifying words. If the mutation is recognized by an existing automaton, then the structure is preserved, in spite of the mutation. Otherwise, this new structure needs to be recognized by a new automaton or it remains meaningless (a semantic null). Subsequently, Sungchul Ji's work[7] on "cell language" generally compared words and biological molecules (RNA, DNA and proteins). This initial concept was refined five years later[8] to include a separation into a DNA language and a protein language, but the definition of a syntactic or even semantic unit remained open.

However, a significant contribution from this work on cell language is the idea that "…human language can be viewed as a transformation of cell language"[7]. If such a theory is pursued further, then the links between semantics and syntax, and the structure–function relationship in proteins, seem less constructed or synthetic than they might at first glance (FIG. 1). And, in fact, it is intriguing how the complexity of computational linguistics has boosted the generation of computational techniques that are used, not only to interpret biological data, but also in a reflexive manner to investigate language and grammar further.

So, if linguistic theories can be applied to biological disciplines like genetics, and maybe even structural biology or evolution, can we establish grammatical rules that will be useful for understanding protein assembly and function? In this article, I discuss the current movements in the field of protein modularity and attempt a provocative look into the future of linguistics-based protein annotation.

Molecular syntax and semantics

Nucleotide codes and amino-acid-based languages are different. Although the genetic information translates directly into the protein, the 'words' that are used in these two individual, but related, languages differ.

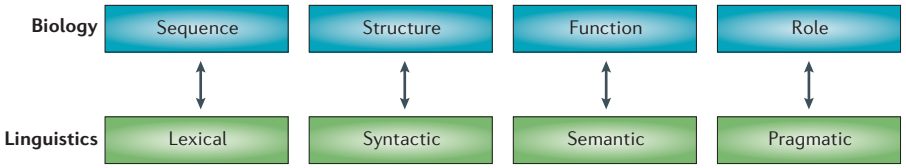

Figure 1 | **Grammatical hierarchies.** The two hierarchies that apply to proteins (top row) and languages (bottom row) express remarkable similarities. Validation of the semantic value of the information that is stored in a sentence or a protein, defines its later pragmatic value or biological role in a larger context. This figure is modified with permission from REF. 17 © (2001) Oxford University Press.

Recently, *in silico* modelling used 'genome semantics' to elucidate the meaning of the genome[13]. In so doing, once more the problem of a suitable definition for syntactic elements became apparent. This work, however, recognized that a semantic code translates sequence into meaning, therefore genome semantics might operate to assign meaning to a regulatory code by means of its function.

David Searls recognized the differences between nucleic-acid and amino-acid codes, and separated nucleic-acid linguistics from protein linguistics[14]. In a series of seminal papers[15–19], Searls identified striking similarities in linguistic complexity between nucleic acids and natural human languages. Most importantly, Searls suggested the significance of the information (or semantic value) that is contained within the linear or folded structure of macromolecules. Moreover, he proposed that the hierarchical order and the rules that govern macromolecular organization relate strongly to syntactic structures in human languages[15,20]. Notably, grammatical ambiguity was elegantly described as reflecting the possibility for alternative (molecular) structures, leaving room for modulations and novel findings.

Different languages. Any discussion about the underlying rules or grammar in a given language or code must start by defining the syntactic building blocks to which this grammar applies. And, exactly at this point, the definition of a unifying general principle for a biological language becomes difficult, as the syntactic and semantic units seem to differ between nucleic-acid and protein languages. In contrast to nucleic-acid sequences, the amino-acid sequences of proteins also define structure and function, implying that protein grammar exhibits a higher level of complexity[6]. But how can we learn to read the biological information that is stored in a finite string of role-forming, three-dimensional structures? And which grammatical rules should we apply?

> "A key theoretical principle for understanding an unknown language is the recognition of syntactic patterns. For proteins, these patterns might be similarities in sequence, or structure, or both."

Modular protein domains

A key theoretical principle for understanding an unknown language is the recognition of syntactic patterns. For proteins, these patterns might be similarities in sequence, or structure, or both. Current genomic and proteomic efforts focus on establishing catalogues of relationships and on constructing family trees on the basis of these seemingly meaningful pattern similarities, implying that a lexicostatistical comparison of entries will allow the functional (semantic) annotation of diverse genomes. But correlating protein function on the basis of similarity might, at present, be too complex to carry out on entire genomes or proteomes, and so a more reductionist approach — using 'modular protein domains' (see BOX 1 for definitions of domains) as the basic syntactic units — seems more fruitful[21,22]. Domains that fold and function autonomously and also convey their inherent function in other proteins are generally referred to as a 'module' (although unfolded domains can also be modular). The term 'module', however, is also used to describe functional interaction hubs that govern networks of interactions[23–28].

Phrases and clauses. Protein language seems more complex than the genetic code, and the features of protein structures seem to more closely resemble complex languages, which make use of conceptual expression patterns, such as Chinese characters or Egyptian hieroglyphs[5,6,29,30]. Considering the domain or module as a word that consists of individual letters (amino acids) is tempting, but biochemistry and structural biology point in another direction. Although the module retains (by biological definition[27]) its autonomous fold and function, it has also been shown that structurally (syntactically) similar modules show functional plasticity (for example, WW domains, calponin homology (CH) domains, pleckstrin homology (PH) domains, and so on; BOX 1; see REF. 22 for a comprehensive overview on modular protein domains). This makes it generally difficult to adopt modular protein domains as words, as they already seem to have a higher semantic complexity.

If, however, we consider a folded domain as a more complex 'phrase' or 'clause', we can more easily accommodate functional plasticity within structurally highly similar protein domains that have little conservation at

Box 1 | **Protein domain definitions**

Domain
Any of the three-dimensional subunits of a protein that, together, make up its tertiary structure, are formed by folding its linear peptide chain, and are variously considered to be the basic units of protein structure, function and evolution.

DH domain
Dbl-homology domains (more recently described as RhoGEF (Rho guanine nucleotide-exchange factor) domains) are ~200-residue modules that are necessary and sufficient for nucleotide exchange activity on small GTPases.

PH domain
Most pleckstrin-homology domains bind various phosphoinositides with different affinities, but they can also mediate specialized protein–protein interactions. It is the eleventh most common protein domain found in the human genome.

CH domain
Calponin-homology domains are versatile, all-helical, 100-residue domains that can mediate binding to cytoskeletal elements like actin filaments, microtubules or intermediate filaments, as well as to Zn^{2+}-finger-like motifs.

WW domain
The smallest known autonomous module that contains 38 residues and recognizes various proline-rich, or proline-containing ligands. The name is derived from two invariantly conserved tryptophan residues that are spaced 20–22 residues apart.

Figure 2 | **The Rosetta Stone.** The Rosetta Stone solved a particularly difficult linguistic problem by providing researchers with the required tool for deciphering hieroglyphs — a form of narrative pictograms. The phrase 'Rosetta Stone' is also used today as a metaphor to refer to anything that provides a key to any process of decryption, translation, or a difficult problem. Image © The British Museum.

the sequence level. Under such conditions, the 'semantic value' (equivalent to the 'information content' or the function(s) that the module or domain performs within a given intramolecular or cellular context) of a protein domain can be modified substantially without changing the length or structure of the clause (its size and basic fold). However, we also have to consider that from a linguistic standpoint, individual words can also have a substantial amount of 'semantic plasticity', indicating that there might be no necessity for viewing modules as phrases. On the other hand, functional plasticity in protein modules does not solely relate to a different behaviour in the context of an altered environment, but rather includes sometimes complex multifactorial behaviours. Indeed, the pioneering work of Tony Pawson on the combinatorial use of protein domains underlines the potential analogies between modular protein elements and linguistic clauses/phrases rather than words[31]. Once again, for human languages, the problem appears to be solved, and each word is treated as an individual unit. By contrast, the problems of precise functional annotation of proteins seem to continue to reside in the lexical and syntactic domains (if one considers that syntax deals with structure of elements that are formed from strings of sequences), while the endpoint of a proper annotation must be the semantic, and eventually pragmatic description of a molecule.

A biological 'Rosetta Stone'? Domains and modules might delineate the authentic syntactic and semantic units in a protein. Are, then, protein modules a biological 'Rosetta Stone' (FIG. 2) that will help us to decipher the first basic rules of grammar in protein linguistics?

Przytycka *et al.*[32,33] elaborated on "protein grammar" by describing four basic grammatical rules that dictate folding of the linear amino-acid sequence into β-protein-domain modules (that consist of β-strands and β-sheets). The recognition of such grammar supports hypotheses that suggest an evolutionary mechanism for the folding of protein domains on the basis of their amino-acid sequence[34–36].

A set of rules for the description of protein–protein-interaction motifs has been developed as a joint effort between scientists that are engaged in modular protein-domain research, and this compilation of standardized nomenclatures is known as the Seefeld Convention[37] (see the Online links box). Again, modular protein domains, rather than the entire protein repertoire, have been used as the basis for this representational grammar. This has become a useful tool for describing functional sites and their cognate recognition modules in a standardized way, using a set of easy-to-follow parameters for proper annotation. In further support of the modular protein-domain concept, the Protein Modules Consortium, a non-profit initiative that unites leading molecular, cellular and computational biologists, as well as geneticists and linguists, who are interested in modular protein domains, has commenced its work (see Online links for details of their recent meeting and other activities).

> "Are, then, protein modules a biological 'Rosetta Stone' that will help us to decipher the first basic rules of grammar in protein linguistics?"

Compositional Semantics

Recognizing the functional domains in a protein and thoroughly validating their individual functions is an essential step towards elucidating the biological function of a given protein. David Searls elaborated on "compositional semantics"[14,17] that relate to the quality and quantity of proteins (or protein modules, in our case). Compositional semantics has a clear mathematical basis — the biological role of a molecule is described by the concerted actions and functions of its individual domains. There are numerous examples in support of this concept, and its applicability can be expanded if we accept that dimerization and multimerization of a molecule is a protein-linguistic operation that increases the semantic value by multiplying its syntactic subunits.

An illustrative example for such a scenario is the formation of the anti-parallel α-actinin homodimer, which is required for the formation of a functional actin-filament crosslinking molecule. Here, each subunit chain contributes one functional actin-binding site, but parallel bundling requires the presence of two oppositely positioned sites. A more dramatic example can be found in the formation of the oncogenically active BCR–ABL fusion protein. The N-terminal fusion of the *ABL* gene to a fraction of the *BCR* gene confers oligomerization and, as a consequence, the activation of the intrinsic tyrosine-kinase activity of the otherwise silent ABL protein[38]. This fusion process generates a novel lexical protein entry with new syntactic properties (FIG. 3), similar to what is observed in evolving, living languages (the development of a politically correct language might be a suitable analogy here).

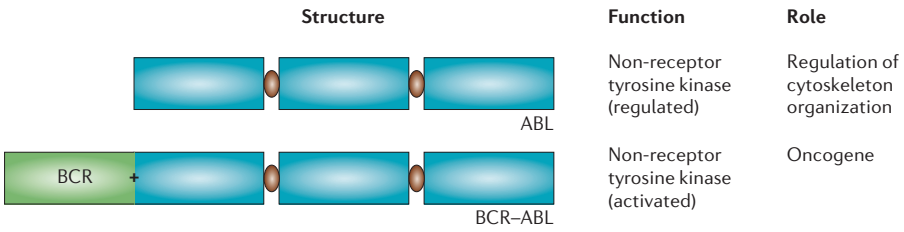

Figure 3 | **Generation of a novel lexical protein entry.** Fusion of parts of the *BCR* gene to the N terminus of the *ABL* gene causes activation of the tyrosine-kinase activity of the ABL protein. The altered behaviour of this novel fusion protein (BCR–ABL) primarily modifies its cellular role (or pragmatic inventory), while the impact on the semantic (functional) level appears modest and the two proteins overlap significantly in the arrangement of their syntactic elements (or protein modules).

Just as proteins can lose their functions when their subcellular localization is altered, protein domains do not always function independently of their position within the protein (similar to syntactic alterations in a clause). Notably, compositional semantics also accounts for the linguistic sensitivity of syntactic manipulation — in our case, the 'biological word order'. As compositional semantics is guided by syntactic analysis, the complexity of positional relationships (similar to dependencies in linguistic analyses) is also supported by this concept. This equates to the protein version of the genome semantics that were discussed earlier in the work of Brendel and Busse[11]. The greater complexity of proteome semantics is easily explained by the three-dimensional structure of a protein and its spatially and temporally regulated functions. Proteins adjoin modules, not only in space, but also in time (by dimerization, folding, protein–protein and protein–lipid interactions, site-unmasking or post-translational modifications) to alter the semantic value of the molecule. Compositional semantics therefore includes the important idea that, not only does the total composition of modules in a protein control its semantic value, but that the position of these modules along the folded amino-acid chain must also be considered. The validity of this theory is illustrated by a recent example in which the output processing sensitivity of a mitogen-activated protein (MAP) kinase pathway was probed by perturbing the scaffold–kinase-assembly mechanisms through the construction of a synthetic 'diverter scaffold'. The engineered scaffold initiated a new signalling cascade with non-natural input–output properties by recruiting a non-native combination of kinases[39].

Intrinsically unfolded proteins

Despite rapid progress in structural genomics, the number of unique folds in the human proteome remains relatively constant at ~100–1,000 (see REF. 21 and references therein). Importantly, only a fraction of all proteins is folded into stable three-dimensional units and, although folded domains represent the most prominent portion of a given protein sequence, the non-folded regions are certainly not superfluous for the function of the protein. Intrinsically unstructured protein (IUP) domains[40] that fold mainly on contact with their ligands might be evaluated as elements that complement 'word stems' or 'roots' (folded domains).

Functional categories and affixes. How can we make protein-linguistic use of IUP regions? If we assume that we have (with the help of structural biology and biochemistry) defined the borders of our module(s), — equivalent to our word, phrase or clause, — then flanking IUP regions could function as mobile functional categories, which include 'minor category words' (for example, 'the', 'is', 'that' and so on) as well as 'prefixes' or 'suffixes'. They do not statically separate one word from the other (this could be the function of non-functional unstructured loops[41]) but, instead, respect the structure of the clause. They might become linguistically activated upon semantic validation of the clause (for example, upon binding to a ligand), which then elicits the function of the 'affix' (binding to a second ligand, a second messenger or a different site on the first ligand) to alter the semantic value (for example, to enhance the binding).

Parameters

The linguist Mark Baker[2] proposes a set of basic parameters that govern all of the currently characterized languages (~6,000) on this planet. He uses these parameters as a tool to explain differences among languages, and draws a surprising comparison between the parameters that define human languages and Mendeleyev's periodic table. Among these parameters, the powerful 'head directionality parameter' is most likely to be applied to a putative protein linguistic grammar. "A verb (module X in the protein equivalent) comes before (or after) the direct interactor (or functional motif/domain in the protein equivalent) to which it is semantically related (to which a functional dependence exists in the protein)." This fundamental principle for universal language grammar could be tremendously helpful for rapid, yet precise, semantic/functional annotation of novel sequences.

There are, in fact, practical biological examples at hand to test if this parameter also has biological value. Two modular protein domains that fulfil the above criteria of 'X' and 'functional dependence' form the double motif DH (Dbl homology)–PH in several molecules (see also the chapter by Lemmon and Keleti in REF. 22). Although both modules exist individually as semantically valuable interaction domains in a large number of proteins (~2,000 in the SMART database), ~600 protein-sequence entries

Glossary

Affix
A meaningful element that cannot stand on its own but it is added to another element.

Automaton
A device that reads input, conventionally from left to right, and either recognizes or generates language.

Clause
A basic unit of grammatical structure that expresses a single thought.

Grammar
The part of a language that is responsible for assembling basic words into larger words, phrases and clauses in systematic ways. For simplicity, grammar may be viewed as a combination of syntax and morphology.

Lexica
The stocks of basic words.

Linguistics
The study of the nature, structure and variation of language (includes the sub-disciplines of morphology, syntax, semantics and pragmatics).

Module
Different languages have different concepts of a module but there are several shared ideas. Modules are similar to objects in an object-orientated language, although a module might contain many procedures and/or functions, which would correspond to many objects. In computer science, modules are described as a portion of a program that carries out a specific function and might be used alone or combined with other modules of the same program.

Phrase
A group of words that appear next to each other or stay together in the arrangement of a sentence and that form a syntactic unit.

Prefix
A meaningful element that cannot stand on its own but it is added to the beginning of another element.

Root
The core of a word, before prefixes and suffixes are attached.

Semantics
The branch of linguistics concerned with the meaning of linguistic expression.

Sentence
A basic unit of a language that expresses a complete thought.

Stem
Prefixes and suffixes attach to a stem in order to form a longer word.

Suffix
A meaningful element that cannot stand on its own but it is added to the end of another element.

Syntax
The branch of linguistics that studies how words are combined to make phrases and sentences.

Word
A freestanding portion of language with a coherent meaning.

> “Every parameter that can be defined will add another piece to the puzzle. But again — caution is advised! Even an appealing collection of metaphors falls short of creating a useful tool.”

from eukaryotes contain both domains. Notably, in over 99% of cases, these domains are present as DH–PH, the semantic value of which mediates the exchange of bound GDP for GTP in small GTPases. According to a biological head directionality parameter, it could be postulated that proteins containing both DH and PH domains arrange these domains in a DH-first manner. If indeed the grammar for all human languages can be defined by a small set of parameters, a similar restriction for the generation of the entire protein complement might be envisaged. It will be interesting to evaluate more of these examples (also using other linguistic parameters as guidelines) to elucidate the generality of this principle. However, this territory should be entered with caution, as a simple, yet fundamental, biochemical reason for such a preferred arrangement will not necessarily argue for the existence of general parameters similar to those introduced by Baker for human languages (discussed in REF. 2).

Concluding remarks

Deciphering the rules underlying a novel language requires access to the semantics (meaning) of the syntactic units, or in simple terms: it is impossible to learn a language without knowing what it means[2]. Words are arbitrary and mobile (or fluid) in their semantic values, and many biologists would readily agree with the extrapolation that constructs that contain a string of arbitrary units create an even more arbitrary compound. But the opposite of this is true in human languages, and the structure of sentences has to follow more stringent conditions than the individual words[2]. Words are also not simply monosemantic strings of characters, and words are not all there is to language. It is instead the combination with the grammar, which defines the principles for their arrangement, and creates a complete language. The grammatical rules are an important part of creativity and provide, rather than withhold, freedom. Grammar is not a simple list of dos and don'ts. A basic grammar that does not cover all aspects and allows for flexibility (and even ambiguity) is still a useful tool. Therefore, it seems important to try to understand the grammar that governs the assembly of modular proteins. Every parameter that can be defined will add another piece to the puzzle. But again — caution is advised! Even an appealing collection of metaphors falls short of creating a useful tool[20].

With this in mind, we might better understand the limitations of genomics and proteomics, which produce predominantly lexical information. The constraints on sentence structure (modular protein assembly) are sufficient for regulating the semantic complexity of a word (the domain of a functional motif) in the context of a sentence (the entire protein). So, a protein domain in isolation might be biologically promiscuous[42], whereas the context of the protein might define its precise meaning[43]. Even in human languages, word order can sometimes seem unimportant, but, for most languages, it does matter[2].

Biochemistry and genetics seem much in need of a system that helps to understand the function(s) of a protein[44]. And conventional single-molecule studies are being rapidly taken over by more comprehensive interactome approaches, which promise to model monocellular and multicellular interaction networks with daunting speed. Marc Vidal[45] predicts the rise of a technology that allows proteome-wide interaction mapping and modelling, and its subcategories — particularly interaction networks based on modular protein-domain interactions[46] — might benefit from a set of rules that can help to define protein function with high fidelity (see Online links box for a selection of interaction databases).

Protein linguistics will probably not fully resolve the problem of ambiguity of functional annotation. But a future perspective is the semantic mapping of proteins through their detailed, yet synthetic, biochemical description[47] coupled with linguistic analyses of their modular protein domains. An encouraging element is the notion that functional complexity is not a matter of proteome size, but rather depends on the protein-domain arrangement[48]. This concept adds more oil to the flame of protein linguistics and compositional semantics — when the smoke has cleared, we all might have become molecular linguists[9]!

Mario Gimona is at the Consorzio Mario Negri Sud, Marie Curie Unit of Actin Cytoskeleton Regulation, Department of Cell Biology and Oncology, Via Nazionale 8A, 66030 Santa Maria Imbaro, Italy.
e-mail: gimona@negrisud.it

doi:10.1038/nrm1785

1. Bogusky, M. S. Biosequence exegesis. *Science* **286**, 453–455 (1999).
2. Baker, M. C. *The atoms of language* (Basic books, New York, 2001).
3. Pesole, G., Attimonelli, M. & Saccone, C. Linguistic approaches to the analysis of sequence information. *Trends Biotechnol*. **12**, 401–408 (1994).
4. Mantegna, R. N *et al.* Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett*. **73**, 3169–3172 (1994).
5. Popov, O., Segal, D. M. & Trifonov, E. N. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems* **38**, 65–74 (1996).
6. Doerfler, W. In search of more complex genetic codes — can linguistics be a guide? *Med. Hypotheses* **9**, 563–579 (1982).
7. Ji, S. Isomorphism between cell and human languages: molecular biological, bioinformatic and linguistic implications. *Biosynthesis* **44**, 17–39 (1997).
8. Ji, S. & Ciobanu, G. Conformon-driven biopolymer shape changes in cell modelling. *Biosystems* **70**, 165–181 (2002).
9. Botstein, D. & Cherry, J. M. Molecular linguistics: extracting information from gene and protein sequences. *Proc. Natl Acad. Sci.* USA **94**, 5506–5507 (1997).
10. Editorial. Folding as grammar. *Nature Struct. Biol.* **9**, 713 (2002).
11. Brendel, V. & Busse, H. G. Genome structure described by formal languages. *Nucleic Acids Res.* **12**, 2561–2568 (1984).
12. Brendel, V., Beckman, J. S. & Trifonov, E. N. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *J. Biomol. Struct. Dyn.* **4**, 11–21 (1986).
13. Werner, E. Genome semantics, *in silico* multicellular systems and the central dogma. *FEBS Lett*. **579**, 1779–1782 (2005).
14. Searls, D. B. Linguistic approaches to biological sequences. *Comput. Appl. Biosci.* **13**, 333–344 (1997).
15. Searls, D. B. in *Artificial Intelligence and Molecular Biology* (ed. Hunter, L.) 47–121 (The MIT Press Classics Series and AAAI press, Cambridge, USA, 1993).
16. Searls, D. B. Using bioinformatics in gene and drug discovery. *Drug Discov. Today* **5**, 135–143 (2000).
17. Searls, D. B. Reading the book of life. *Bioinformatics*, **17**, 579–580 (2001).
18. Searls, D. B. The language of genes. *Nature*, **420**, 211–217 (2002).
19. Searls, D. B. Trees of life and of language, *Nature* **426**, 391–392 (2003).
20. Dong, S. & Searls, D. B. Gene structure prediction by linguistic methods. *Genomics* **23**, 540–551 (1994).
21. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
22. *Modular Protein Domains*. (eds Cesareni,G., Gimona, M., Sudol, M. & Yaffe, M.) (WILEY-VCH, Weinheim, 2004).
23. Papin, J. A., Hunter, T., Palsson, B. O. & Subramaniam, S. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Rev. Mol. Cell Biol*. **6**, 99–111 (2005).
24. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
25. Han, J.-D. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93 (2004).
26. Wuchty, S. Scale-free behaviour in protein domain networks. *Mol. Biol. Evol.* **18**, 1694–1702 (2001).
27. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
28. Wuchty, S., Oltvai, Z. N. & Barabasi, A.-L. Evolutionary conservation of motif constituents in the yeast interaction network. *Nature Genet.* **35**, 176–179 (2003).
29. Pietrokovski, S, Hishon, J. & Trifonov, E. N. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *J. Biomol. Struct.* **7**, 1251–1268 (1990).
30. Pietrokovski, S & Trifonov, E. N. Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. *Gene* **122**, 129–137 (1992).
31. Pawson, T. Protein modules and signalling networks. *Nature* **373**, 573–580 (1995).

32. Przytycka, T., Aurora, R. & Rose, G. D. A protein taxonomy based on secondary structure. *Nature Struct. Biol.* **6**, 672–682 (1999).
33. Przytycka, T., Srinivasan, R. & Rose, G. D. Recursive domains in proteins. *Prot. Sci.* **11**, 409–417 (2002).
34. Sim, J., Kim, S. Y. & Lee, J. PPRODO: prediction of protein domain boundaries using neural networks. *Proteins* **59**, 627–632 (2005).
35. Sonnhammer, E. L. L. & Kahn, D. Modular arrangement of proteins as inferred from analysis of homology. *Prot. Sci.* **3**, 482–492 (1994).
36. Galzitskaya, O. V. & Melnik, B. S. Prediction of protein domain boundaries from sequence alone. *Prot. Sci.* **12**, 696–701 (2003).
37. Aasland, R. *et al.* Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.* **513**, 141–144 (2002).
38. Arlinghaus, R. B. Bcr: a negative regulator of the Bcr–Abl oncoprotein in leukemia. *Oncogene* **21**, 8560–8567 (2002).
39. Park, S.-H., Zarrinpar, A. & Lim, W. A. Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. *Science* **299**, 1061–1064 (2003).
40. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
41. George, R. A. & Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Prot. Eng.* **15**, 871–879 (2002).
42. Pawson, T. Specificity in signal transduction: from phosphotyrosine–SH2 domain interactions to complex cellular systems. *Cell* **116**, 191–203 (2004).
43. Farooq, A., Sudol, M. & Zhou, M.-M. Two is better than one: structure function and mechanism of tandem domains. *Nova Publications* (in the press).
44. Benner, S. A. & Gaucher, E. A. Evolution, language and analogy in functional genomics. *Trends Genet.* **17**, 414–418 (2001).
45. Vidal, M. Interactome modelling *FEBS Lett.* **579**, 1834–1838 (2005).
46. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
47. Sudol, M. From src homology modules to other signalling domains: proposal of the "Protein Recognition Code". *Oncogene* **17**, 1469–1474 (1998).
48. Wuchty, S. & Almaas, E. Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.* **5**, 24 (2005).

Acknowledgements

I wish to thank M. C. Baker and M. Sudol for critically commenting on this manuscript, and the members of the Protein Modules Consortium for inspiring discussions. The author is supported by a Marie Curie Excellence Grant of the Framework Program 6 of the European Union.

Competing interests statement

The author declares no competing financial interests.

DATABASES

The following terms in this article are linked online to:
Artificial Intelligence and Molecular Biology (electronic text (PDF) of the out-of-print book): http://www.aaai.org/Library/Hunter/hunter.html
Cytoscape: www.cytoscape.org
FEBS workshop on Modular Protein Domains: from functional plasticity to protein linguistics (official meeting web site): http://www.negrisud.it/en/congres/Seefeld05/
Protein Modules Consortium: www.proteinmodules.org
The BIND interaction database: http://bind.ca
The DIMA domain interaction map: http://mips.gsf.de/genre/proj/dima/links.html
The InterPro database: http://www.ebi.ac.uk/interpro/
The MINT database: http://cbm.bio.uniroma2.it/mint/index.html
The Seefeld Convention: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=11911894&query_hl=14

FURTHER INFORMATION

Mario Gimona's laboratory: http://www.negrissud.it/en/dcbo
Access to this interactive links box is free online.