Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 328 (2005) 882–899
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Review / Revue

Protein variety and functional diversity: Swiss-Prot annotation in its biological context

Brigitte Boeckmann [a,*], Marie-Claude Blatter [a], Livia Famiglietti [a], Ursula Hinz [a], Lydie Lane [a], Bernd Roechert [a], Amos Bairoch [a,b]

[a] *Swiss Institute of Bioinformatics, Centre médical universitaire, 1, rue Michel-Servet, 1211 Genève 4, Switzerland*
[b] *Department of Structural Biology and Bioinformatics, Centre médical universitaire, 1, rue Michel-Servet, 1211 Genève 4, Switzerland*

Received 13 May 2005; accepted after revision 5 June 2005

Available online 28 July 2005

Presented by Stuart Edelstein

Abstract

We all know that the dogma 'one gene, one protein' is obsolete. A functional protein and, likewise, a protein's ultimate function depend not only on the underlying genetic information but also on the ongoing conditions of the cellular system. Frequently the transcript, like the polypeptide, is processed in multiple ways, but only one or a few out of a multitude of possible variants are produced at a time. An overview on processes that can lead to sequence variety and structural diversity in eukaryotes is given. The UniProtKB/Swiss-Prot protein knowledgebase provides a wealth of information regarding protein variety, function and associated disorders. Examples for such annotation are shown and further ones are available at http://www.expasy.org/sprot/tutorial/examples_CRB. ***To cite this article: B. Boeckmann et al., C. R. Biologies 328 (2005).***

Résumé

Un gène, plusieurs protéines : l'annotation de Swiss-Prot dans le contexte biologique. Il est maintenant évident pour tout le monde que le dogme « un gène, une protéine » est obsolète. Au cours de la synthèse d'une protéine fonctionnelle, le transcrit et la chaîne polypeptidique peuvent être modifiés de multiples façons. Ces modifications ont une incidence directe sur la fonction biologique de la protéine et dépendent non seulement de l'information génétique, mais également des conditions dans lesquelles se trouve la cellule : un nombre limité d'isoformes protéiques est produit dans une cellule donnée, à un moment précis. Cet article dresse un bref inventaire des processus biologiques impliqués dans la formation de protéines différentes à partir d'un même gène chez les eucaryotes, ainsi qu'une description des diversités structurelle et fonctionnelle qui en découlent. La banque de connaissances sur les protéines UniProtKB/Swiss-Prot est particulièrement riche en informations décrivant l'origine des différences entre les séquences de protéines dérivées d'un même gène, les modifications post-traductionnelles, ainsi que les conséquences de cette variabilité sur leur(s) fonction(s) et, le cas échéant, les maladies associées. De nombreux exemples

* Corresponding author.
E-mail address: brigitte.boeckmann@isb-sib.ch (B. Boeckmann).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2005.06.001

d'annotation sont décrits et d'autres sont disponibles sur le site http://www.expasy.org/sprot/tutorial/examples_CRB. ***Pour citer cet article : B. Boeckmann et al., C. R. Biologies 328 (2005).***

Keywords: Protein database; Annotation; Protein synthesis; Sequence variety; Post-translational modification; Protein–protein interaction; Disease

Mots-clés : Banque de données sur les protéines ; Annotation ; Synthèse protéique ; Diversité des séquences ; Modifications post-traductionnelles ; Interaction protéine–protéine ; Maladie

1. Introduction

Despite an abundant biodiversity, all living beings are based on a similar cellular system, which is run by a population of self-organized molecules: proteins. Proteins catalyze, regulate and control most procedures that occur in a cell for the benefits of the whole organism. If we wish to understand how living systems work, it is important to understand how proteins function. The way life evolved facilitates this task in that we can compare not only organisms but also physiological processes and their components. Consequently, conclusions can be drawn from the findings and applied from one system to another. In the past decades, numerous methods – e.g., sequence comparisons, protein family prediction, detection of functional domains, conserved domains and altered amino-acid positions within these regions, motif searches, structure prediction or phylogenetic studies – and databases have been developed for the efficient prediction of a protein's function. However, such methods require the 'correct' amino-acid sequence as input; the retrieval of such input is still a challenging task [1]. In eukaryotes especially, the formation of the nascent amino-acid sequence implies the possible creation of a huge number of isoforms. Without further experimental evidence it is impossible to predict the existing and biologically relevant proteins from the total of all possible variants. The same is true for most of the other alterations in a protein's structure. What is more, there is still a long way to go to understand how polypeptides interact with each other in order to accomplish a specific task. In this respect, the study of inheritable diseases can be very informative and demonstrate how the smallest of deviations from a protein's structure can lead to its dysfunction and be the cause of severe disorders.

This article gives an overview on cellular processes underlying sequence variety and structural diversity. We have chosen to focus on eukaryotes to narrow down our discussion. The UniProtKB/Swiss-Prot protein knowledgebase [2,3] aims to record all protein variations and their functional impact. Throughout the text, examples of corresponding Swiss-Prot annotation are given and the reader is encouraged to look at further examples when the primary accession number is indicated (e.g. P12345). Complete Swiss-Prot examples are provided at http://www.expasy.org/sprot/tutorial/examples_CRB. Swiss-Prot entries can also be retrieved from the ExPASy server [4] by building the URL, e.g., http://www.expasy.org/uniprot/P12345.txt for the raw format, http://www.expasy.org/uniprot/P12345 for the NiceProt format, or by entering the accession number in the quick search at http://www.expasy.org/ (NiceProt format). Further examples of Swiss-Prot entries can be found via the relevant keywords. Details on the format of Swiss-Prot entries are provided in the user manual at http://www.expasy.org/sprot/userman.html.

2. Formation of the nascent amino-acid sequence

The vast majority of eukaryotic proteins are nuclear-encoded, as are most of the proteins which are the products of DNA-containing organelles. Mitochondria and plastids generate only a small fraction of their own proteome [5–10]. Regulatory mechanisms during protein synthesis can influence the concentration, destination, sequence variety, structural diversity and thus the functional options of the resulting protein. Out of a broad variety of possible mRNAs and protein sequences that can be generated from a single gene (Fig. 1), only one or a few are created at a time, dependent on the type of tissue or the stage of development.

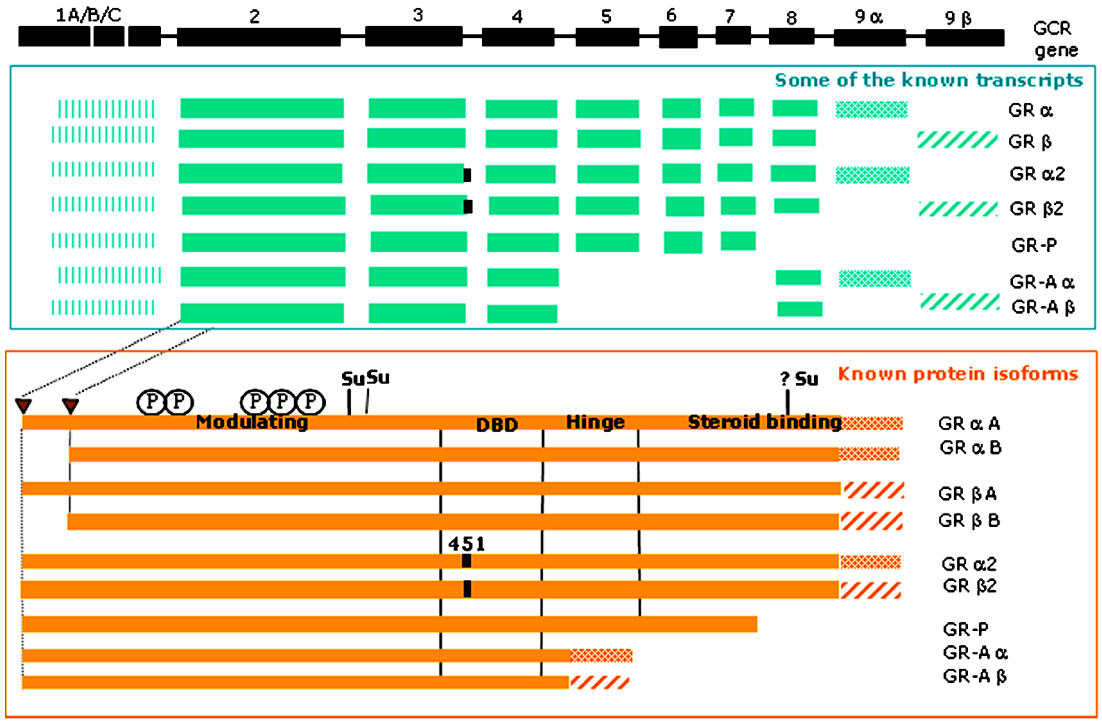

Fig. 1. Known sequence isoforms of the human glucocorticoid receptor (GCR) (P04150). The formation of the protein isoforms is based on alternative splicing events, including partial intron retention at the 3′ end of exon 3 (amino-acid position 451 of the relevant GCR isoforms) and alternative translation initiation (A- and B-type isoforms; it is not known yet if this occurs for each mRNA). Noteworthy, exons 9a and 9b are mutually exclusive (giving rise to isoforms alpha and beta, respectively). Further transcripts are created, which differ in their 5′ untranslated region due to alternative promoter usage (exon 1A/B/C) and alternative splicing within exon 1A. The PTMs (P: phosphorylation; Su: SUMOylation) and the different domains (DBD: nuclear hormone receptor DNA binding domain) of the GCR are indicated.

The initial step of transcription already gives rise to the production of different transcripts provided that a given gene expression is controlled by more than one promoter. Alternative promoter usage (Q9S7T7) can influence transcription initiation, mRNA stability as well as translation efficiency, thus causing the formation of distinct protein isoforms [11].

The next step – the processing of the primary transcript – is essential for pre-mRNA protection, mRNA export from the nucleus to the cytosol and for efficient translation and translation regulation [12]. In higher eukaryotes, most transcripts contain introns [13] that are removed co- and post-transcriptionally by way of RNA splicing [14,15]. The number of introns is highly variable from one gene to another, within the same organism and between species. In *S. cerevisiae*, for example, only about 255 of the 6200 expected genes contain introns (an average of one to two [16]) whereas most human genes are thought to contain at least one intron but often many more. The human titin gene contains 363 exons, but it has to be noted that it encodes for an unusually large protein of 38 138 residues [17]. Higher eukaryotes generally perform alternative splicing [18], which is a powerful mechanism for the creation of sequence variety. In humans, it has been estimated that 40 to 60% of the genes are alternatively spliced [19,20]. The production of an extensive variety of mRNA isoforms is achieved through alternative exon splicing, extension of the 5′ or 3′ boundaries of exons or the retention of introns [21,22]. The resulting mRNAs can differ in their regulatory regions and coding region. In the latter case, changes can cause the substitution, insertion or deletion of one (P36542) or more amino acids in the protein isoforms (P30429). Polypeptides can also differ in the composition of their domains (see Section 4). Other modifications can affect the location of stop codons and thus generate truncated or extended

isoforms. In the case of the Drosophila gene 'dscam', the 95 exons of the transcript could potentially give rise to 38 016 different protein isoforms [23]. An example for the formation of differing isoforms would be the leptin receptor gene (LEPR) (O15243), whose mRNA only shares the two 5′ untranslated exons with the canonical isoform (P48357). Exon order is normally conserved subsequent to alternative splicing but there are examples where exons are joined in an order different from that in the genome [24]. RNA splicing can also combine exons that originate from more than one gene (trans-splicing). An example of a chimeric mRNA is the human cytochrome P450 3A, which is made up of exons from the CYP3A43 and CYP3A4 or CYP3A5 genes (Q9HB55) [25].

The genetic information of a transcript can be further changed via mRNA editing thus possibly giving rise to functionally differing proteins [26]. In higher eukaryotes, the bases uracil and inosine are produced by the hydrolytic deamination of cytosine and adenosine, respectively [27], which could lead to the substitution of an amino acid in the polypeptide. The generation or modification of a stop codon within a reading frame of the mRNA creates an isoform that differs in its C-terminal portion (P04114). RNA editing can also modify the reading frame when bases within a coding region are deleted or inserted, as it has been reported for mitochondria in primitive eukaryotes [28] (Q07434).

After maturation, the mRNA is exported from the nucleus to the cytosol via the nuclear pores by a mechanism which has been conserved from yeast to humans [29]. Translation starts immediately at the cellular translation machinery [30,31]. For the great majority of mRNAs (90%), the initiation site is the first cap-proximal initiator codon located in the appropriate sequence context [30,32]. The initiator tRNA contains an anticodon which is specific for AUG (Met), rarely CUG (Leu), UUG (Leu) or GUG (Val) [33]. Whatever codon is used, methionine seems to be the first amino acid in the nascent polypeptides, except for the CUG-initiated MHC class I bound peptides, which can start with leucine [34]. The translation machinery can make use of alternative initiation codons. In particular, the usage of non-AUG codons at alternative initiation sites has been extensively observed in regulatory proteins such as proto-oncogenes, transcription factors, kinases and growth factors [35]. If the alternative translation initiation codon lies in the same reading frame, polypeptides which differ in their N-terminus are generated. If the long isoform includes an N-terminal transfer signal, which is missing in the shorter isoforms, the generated polypeptides will have distinct subcellular destinations (P08037). If the alternative initiation codon is not located in the same reading frame, the resulting polypeptide can be completely different, as shown for the Sendai virus P/V/C gene (P04862).

In the process of translation, non-standard decoding mechanisms can further manipulate the genetic information. Such mechanisms are known as 'recoding'. Recoding events seem to be rare but probably exist in many organisms [36,37]. In a process named 'programmed ribosomal frameshifting', the translating ribosomes slide one base backward (−1 frameshifting) or forward (+1 frameshifting) at a specific site on the mRNA [36,38,39] (O95190). 'Stop codon readthrough' is a mechanism by which the meaning of the stop codon is redefined. Well-studied examples are the incorporation of the amino acids selenocysteine (UGA) [40] (P24183) and pyrrolysine (UAG) [41] (O30642). According to our current knowledge, the latter has only been observed in prokaryotes [42].

In Swiss-Prot, all protein variants of a gene are generally described in a single database entry. Cellular mechanisms that lead to an amino-acid sequence differing from the one expected by standard translation of the nucleotide sequence are indicated. Examples of such annotation are given in Table 1. Unlike many other annotations, information on sequence variety is currently not transferred to the corresponding entries of closely related eukaryotes as the level of conservation of such processes in the course of evolution is not yet well defined.

3. Protein sorting and associated sequence modifications

Once synthesized, proteins are usually further processed even if they remain in the same cellular compartment. The initiator methionine of many cytosolic proteins is co-translationally cleaved. This is frequently followed by the acetylation of the new N-terminus [43] (P07327). Such proteins fold immediately and are then transported to their final destination.

Q9S7T7: Alternative promoter usage

```
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event = Alternative promoter;
CC         Comment = 2 isoforms, 2 (shown here)
CC         and 1, are produced by use of
CC         alternative promoters;
KW   Alternative promoter usage.
FT   VARSPLIC   1    16     MLRLESLIVTVWGPAT ->
FT                          MFYSHQ (in isoform 1).
FT                          /FTId = VSP_007492.
```

P30429: Alternative splicing

```
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event = Alternative splicing; Named isoforms = 2;
CC       Name = b; Synonyms = Long;
CC         IsoId = P30429-1; Sequence = Displayed;
CC         Note = Minor transcript;
CC       Name = a; Synonyms = Short;
CC         IsoId = P30429-2; Sequence = VSP_013199;
CC         Note = Major transcript;
KW   Alternative splicing.
FT   VARSPLIC 212  234     ARVVSDTDDSHSITDFINRVLSR
FT                         -> K (in isoform a).
FT                         /FTId = VSP_013199.
```

P04114: mRNA editing

```
CC   -!- RNA EDITING: Modified_positions = 2180;
CC       Note = The stop codon (UAA) at position 2180
CC       is created by RNA editing. Apo B-48, derived
CC       from the fully edited RNA, is produced only
CC       in the intestine and is found in chylomicrons.
CC       Apo B-48 is a shortened form of apo B-100 which
CC       lacks the LDL-receptor region. The unedited
CC       version (apo B-100) is produced by the liver
CC       and is found in the VLDL and LDL.
KW   RNA editing;
FT   CHAIN     28 4563    Apolipoprotein B-100.
FT   CHAIN     28 2179    Apolipoprotein B-48.
```

P19970: Alternative translation initiation

```
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event = Alternative initiation;
CC         Comment = 2 isoforms, Long (shown here) and
CC         Short, are produced by alternative initiation.
CC         The isoform Long maintains rhythms at high
CC         temperature (30 degrees Celsius), while the
CC         isoform Short maintains rhythms at lower
CC         temperature (18 degrees Celsius);
KW   Alternative initiation;
FT   CHAIN      1  989       Frequency clock protein,
FT                          isoform Long.
FT   CHAIN    100  989       Frequency clock protein,
FT                          isoform Short.
FT   INIT_MET 100  100       For isoform Short.
```

O95190: Ribosomal frameshifting

```
CC   -!- MISCELLANEOUS: A ribosomal frameshift occurs
CC       between the codons for Ser-32 and Asp-33. An
CC       autoregulatory mechanism enables modulation
CC       of frameshifting according to the cellular
CC       concentration of polyamines.
KW   Ribosomal frameshift.
```

P24183: Stop codon readthrough 1

```
KW   Selenocysteine.
FT   SE_CYS      196    196
```

O30642: Stop codon readthrough 2

```
KW   Pyrrolysine.
FT   MOD_RES     201    201       Pyrrolysine.
```

Translocation to another subcellular compartment requires both protein transfer across at least one membrane and the existence of at least one translocation signal specific to the relevant trafficking mechanism. Major membrane transfer complexes are indicated in Table 2. During transport, nascent proteins are usually associated with chaperones that keep the former in a partly unfolded, soluble conformation whilst protecting them from abnormal folding and aggregation [60]. The chaperones and possibly other factors escort the polypeptide to the relevant membrane receptors of the translocation complex. Protein transfer signals can be proteolytically cleaved during their passage through the membrane (see Table 2), and polypeptides that have to cross several membranes to reach their destination might be cleaved more than once.

There are proteins – such as importins – that act in distinct subcellular compartments and have to cross the membrane in both directions. Such proteins make use of the bi-directional nuclear pore complex [44,61], whereas other membrane-transfer complexes support either the import or the export of polypeptides. Secretory routing via the endoplasmic reticulum (ER) [62,63] is accomplished by proteins which are designated not only for secretion but also for their incorporation into the organelles which are an integral part of the secretion path: the ER, the Golgi apparatus and the lysosomes. Such proteins generally have an N-terminal signal peptide which is removed during membrane transfer.

Various mechanisms have been described for the incorporation of proteins into different types of membranes (Table 2). Common examples are the type-I and type-II membrane proteins. The signal peptide

Table 1
Swiss-Prot annotation of cellular mechanisms leading to an amino-acid sequence different from the one expected by standard translation of the nucleotide sequence. Alternative promoter usage, alternative splicing and alternative initiation events are described in 'CC ALTERNATIVE PRODUCTS', with corresponding information on the sequence changes described in the FT lines ('FT VARSPLIC', 'FT CHAIN' and 'FT INIT_MET', resp.). RNA editing data are annotated in the 'CC RNA EDITING' line and, depending on the event, also in the 'FT CHAIN' or 'FT VARIANT' lines. Ribosomal frameshifting is annotated in the 'CC MISCELLANEOUS' line. A special FT key, 'SE_CYS', exists for stop codon readthrough with selenocystein incorporation. Currently, pyrrolysine integration is still indicated under the feature key 'MOD_RES'. Each of these events has a corresponding keyword (KW line)

Table 2
Well-studied membrane-transfer mechanisms. (1) N-terminal transfer signal peptides are typically cleaved but exceptions have been found (P05120). In some cases, the uncleaved signal peptide confers important functional properties to the protein [56] (P27170). Abbreviations: TOM = translocase of outer mitochondrial membrane, SAM = sorting and assembly machinery, TIM = translocase of inner mitochondrial membrane, TIM23 = presequence translocase, TIM22 = carrier translocase, PAM = presequence translocase-associated motor, TOC = translocon at the outer membrane of chloroplasts, TIC = translocon at the inner membrane of chloroplasts, TAT = twin-arginine translocation, NPC = nuclear pore complex, PEX = peroxin

Translocation (from/to)	Transfer complex	Cleaved signal	References
Nucleus			
Cytosol/nucleus (bi-directional)	NPC	No	[44]
Mitochondrion			
Cytosol			
→ outer membrane	TOM/SAM	No	[45]
→ inner membrane	TOM–TIM22/TIM54	No	[46]
→ matrix	TOM–TIM23/TIM17	Yes	
→ intermembrane space	TOM/Mia40	No	[47]
Matrix			
→ inner membrane	Oxa1	No	[48]
	Oxa2	No	[49]
Chloroplast			
Cytosol			
→ outer membrane	Spontaneous?	No	[50,51]
→ inner membrane	TOC	Yes (1)	
→ stroma	TOC/TIC	Yes	
Stroma			
→ thylakoid membr.	Spontaneous	Yes	[52]
	ALB3	Yes	
	SEC		
→ thylakoid lumen	TAT	Yes	[53]
	SEC	Yes	
Endoplasmic reticulum			
Cytosol			
→ extracellular	Sec61	Yes (1)	[54]
→ type I membr. p.	Sec61	Yes (1)	[55]
→ type II membr. p.	Sec61	No	[56]
Peroxisome			
Cytosol			
→ membrane	PEX3/PEX16/PEX19	No	[57]
→ matrix	Unknown,	No	[58]
	(> 20 PEX)		[59]

of type-I membrane proteins is typically cleaved before the N-terminal of the polypeptide is transferred into the ER lumen. The C-terminal part of the protein remains in the cytosol. As for type-II membrane proteins, the exact contrary occurs: their C-terminal part is transferred into the ER lumen whilst the N-terminal part remains in the cytosol and the internal signal anchor functions as a transmembrane region. Polypeptides without a hydrophobic segment can still be anchored to intracellular or plasma membranes by the covalent attachment of a lipophil group such as glycosylphosphatidylinositol (GPI) (P04058), isoprenoid (P40855), myristate (P62330) or palmitate (O43687). As an example, prenylated proteins – that are estimated to represent 0.5% of all intracellular proteins – have in fact been found on the cytoplasmic surface of plasma membranes, peroxisomal membranes and nuclear membranes [64]. Transport mechanisms are generally well conserved throughout the kingdoms of life, and mitochondria as well as plastids contain translocation systems in their membranes, which point back to their bacterial origin [49,51,53,65]. Dysfunction in

transport systems has been shown to be associated with a variety of diseases [66]. An example would be a dysfunction in the nuclear-cytoplasmic transport system, which has been found to give rise to distinct types of cancer [67].

Various aspects of protein sorting and associated protein processing are reported in Swiss-Prot. Extracts from Swiss-Prot entries show relevant annotation in Table 3.

P07327: Cytosolic protein

```
CC   -!- SUBCELLULAR LOCATION: Cytoplasmic.
KW   Acetylation.
FT   INIT_MET      0      0
FT   MOD_RES       1      1         N-acetylserine.
```

Q9NRA8: Nuclear and cytosolic protein

```
CC   -!- SUBCELLULAR LOCATION: Cytoplasmic; predominantly.
CC       Shuttles between the nucleus and the cytoplasm
CC       in a CRM1-dependent manner.
KW   Nuclear protein.
FT   MOTIF       195    211       Nuclear localization signal.
FT   MOTIF       438    447       Nuclear export signal.
FT   MOTIF       613    638       Nuclear export signal.
FT   MUTAGEN     195    196       RR->NS: Abolishes the
FT                                nuclear localization.
```

P28330: Mitochondrial matrix protein

```
DE   … mitochondrial precursor …
CC   -!- SUBCELLULAR LOCATION: Mitochondrial matrix.
KW   Mitochondrion; Transit peptide.
FT   TRANSIT       1     30       Mitochondrion.
FT   CHAIN        31    430       Acyl-CoA dehydrogenase,
FT                                long-chain specific.
```

O82660: Chloroplast thylakoid lumen protein

```
DE   … chloroplast precursor …
CC   -!- SUBCELLULAR LOCATION: Chloroplast; within the
CC       thylakoid lumen but attached to the membrane.
CC       Restricted to the stromal lamelae.
KW   Chloroplast; Membrane; Thylakoid; Transit peptide.
FT   TRANSIT       1     53       Chloroplast (Potential).
FT   TRANSIT      54     78       Thylakoid (Potential).
FT   CHAIN        79    403       Photosystem II stability/
FT                                assembly factor HCF136.
```

P02724: Type I membrane protein

```
DE   … precursor …
CC   -!- SUBCELLULAR LOCATION: Type I membrane protein.
KW   Signal; Transmembrane.
FT   SIGNAL        1     19
FT   CHAIN        20    150       Glycophorin A.
FT   TOPO_DOM     20     91       Extracellular.
FT   TRANSMEM     92    114
FT   TOPO_DOM    115    150       Cytoplasmic.
```

P10852: Type II membrane protein

```
CC   -!- SUBCELLULAR LOCATION: Type II membrane protein.
KW   Signal-anchor; Transmembrane.
FT   TOPO_DOM      1     75       Cytoplasmic (Potential).
FT   TRANSMEM     76     99       Signal-anchor for type II
FT                                Membrane protein.
FT   TOPO_DOM    100    526       Extracellular (Potential).
```

P80035: Secreted protein

```
DE   … precursor …
CC   -!- SUBCELLULAR LOCATION: Secreted.
KW   Signal.
FT   SIGNAL        1     19
FT   CHAIN        20    398       Gastric triacylglycerol
FT                                lipase.
```

4. Protein folding and structure

Polypeptides fold spontaneously and some of them can attain their native conformation unaided in an extremely short lapse of time [68]. Many other proteins, however, require the assistance of chaperones and possibly additional folding factors and enzymes such as protein disulfide-isomerases or prolyl isomerases, which protect the polypeptides from aggregating and help them reach their native state [69,70]. Both the folding pathway and the three-dimensional structure are dictated by the amino-acid sequence of a polypeptide. Protein folding is driven by the free energy of conformation that is gained by going to a stable, native state [68]. Helices and beta-strands are rich in hydrogen bonds and therefore energetically favourable structures, which is why they form spontaneously in most unfolded proteins. Contacts between these elements give rise to local, native-like conformations and their nucleation facilitates achieving the final structure. This process implies a stochastic approach since the process of protein folding goes through a number of folding intermediates [68–72]. The native state is stabilized by favourable interactions both at the surface and inside the folded polypeptide chain, and involves direct contacts between amino-acid residues as well as contacts with water molecules or ions. Disulfide bonds strongly increase the stability of a protein; this is particularly important for proteins consisting of short chains, which have been derived from larger precursors [68]. Even though the protein structure is a stable construct, it can contain regions of conforma-

Table 3
Swiss-Prot annotation relevant to protein sorting and associated protein modifications. The location of the functional protein is annotated in the CC line topic ‘SUBCELLULAR LOCATION’. Cleaved transfer peptides are indicated in the feature table by the feature keys ‘SIGNAL’ for the signal peptide essential in the secretory pathway and ‘TRANSIT’ for the transit peptide which promotes transfer to mitochondria or chloroplasts. The extent of the mature protein is indicated in the feature key ‘CHAIN’ and the protein name given in the DE line is followed by the term ‘precursor’. The FT key ‘TRANSMEM’ is used to annotate both, transmembrane domains and signalling anchors; ‘MOTIF’ is used for short stretches of transfer signals which are not removed. Keywords (KW) refer to the subcellular location, the type of signal, and sometimes the routing paths

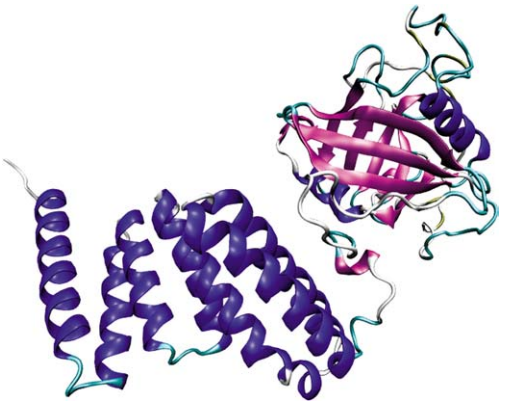

Fig. 2. Ribbon plot of bovine 40-kDa peptidyl-prolyl *cis-trans* isomerase, generated from PDB 1IHG. On the right side the catalytic PPIase domain with a high content of beta-sheet, on the left side a second domain containing three tetratrico peptide repeats (TPR) that mediate protein–protein interactions. The TPR repeat is a degenerate repeat motif of ca. 34 amino acids, each containing 2 antiparallel helices. The minimal functional unit seems to be a structure containing three such repeats.

tional flexibility and many enzymes undergo allosteric transitions (P00489).

Protein structures are strongly conserved in evolution and are the basis for protein classification [73, 74]. Most globular proteins contain both alpha helices and beta-strands; other classes of proteins have either only an alpha-helical structure or purely a beta-stranded one, such as aquaporin-CHIP (P29972) or the beta-barrel porin ompF (P02931), respectively. In the native structure, helices and beta-strands are often connected by beta-turns or mobile loops, thus forming structural motifs. Repeats consist of small structural elements, each of which is too short to be stable but multiple consecutive copies stabilize each other and form typical super-structures, such as the propeller-like structure consisting of Kelch repeats or the arch-like shape made up of leucine-rich (LRR) repeats [75]. Domains are stable, independent folding units, with a characteristic secondary structure topology. On an average, the smallest domains are about 35 amino acids long, but large domains can consist of several hundred amino acids. The average size of a domain is about 160 residues [76]. Usually, short domains such as the zinc-finger and EGF-like domains are stabilized by metal ligands or disulfide bonds. In a given protein, domains often have specific functions, as exemplified by the 40-kDa peptidyl-prolyl *cis-trans* isomerase (P26882) (Fig. 2). This enzyme contains two structural domains: a catalytic PPI domain and a region consisting of three TPR repeats that mediates interactions with heat-shock proteins.

Table 4
Swiss-Prot annotation relevant to the protein structure, extracted from the entry P26882. Experimental information is indicated in the RP line of the relevant reference. Similarity to domains or a protein family is given in the comment (CC) line topic 'SIMILARITY'. Cross-references to structure-related databases are recorded in the DR lines. The keyword '3D-structure' (KW) is added to an entry whenever the 3D-structure of the protein or part of it has been resolved experimentally. The boundaries of repeats, domains, helices, beta-strands and turns are annotated in the feature table (FT)

```
RP   X-RAY CRYSTALLOGRAPHY (1.8 Å).
CC   -!- SIMILARITY: Contains 1 PPIase cyclophilin-type domain.
CC   -!- SIMILARITY: Contains 3 TPR repeats.
DR   PDB; 1IHG; X-ray; A = 1-369.
DR   PDB; 1IIP; X-ray; A = 1-369.
DR   InterPro; IPR002130; CSA_PPIase.
DR   InterPro; IPR001440; TPR.
DR   InterPro; IPR008941; TPR-like.
DR   Pfam; PF00160; Pro_isomerase; 1.
DR   Pfam; PF00515; TPR; 3.
DR   PRINTS; PR00153; CSAPPISMRASE.
DR   SMART; SM00028; TPR; 3.
DR   PROSITE; PS00170; CSA_PPIASE_1; 1.
DR   PROSITE; PS50072; CSA_PPIASE_2; 1.
DR   PROSITE; PS50005; TPR; 3.
DR   PROSITE; PS50293; TPR_REGION; 2.
KW   3D-structure; Direct protein sequencing;
KW   Repeat; TPR repeat.
FT   DOMAIN       18    182       PPIase cyclophilin-type.
FT   REPEAT      222    255       TPR 1.
FT   REPEAT      272    305       TPR 2.
FT   REPEAT      306    339       TPR 3.
FT   STRAND       27     36
FT   TURN         38     40
FT   HELIX        42     53
```

The proteome of higher eukaryotes contains a high proportion of multi-domain proteins and a multitude of different domain combinations, which are hypothesized to have arisen by gene duplication and exon shuffling events. These mechanisms are thought to be a driving force for the rapid evolution of new proteins and more complex organisms. [77,78]. Alternative exon usage can change the number and type of domains present in the protein and has profound effects on a protein's function, such as its capability to interact with other proteins and ligands [79] (Fig. 1).

In a Swiss-Prot entry, the extent of structural elements is recorded in the feature table (Table 4). Cross-references to structure-related databases facilitate access to complementary information.

5. Function-related protein modifications

The majority of protein modifications occur post-translationally, i.e. once the protein has undergone folding, and is typically catalyzed by specific enzymes found in the ER, the Golgi apparatus, the cytoplasm or the nucleus. In the literature – as in Swiss-Prot – the term PTM (Post Translational Modification) is often used in a rather general sense and includes both co- and post-translational modifications. Protein alterations can be relevant to the transport of the nascent protein, to protein folding, and to the activity and function of the native protein. This section is addressed to function-related protein modifications.

Various native proteins, including structural proteins, hormones, neuropeptides and secreted enzymes, are cleaved to achieve their mature form. As an example, receptors of the notch signalling pathway (P46531) are cleaved after ligand-binding, to release the cytoplasmic domain which then acts as a transcription factor in the nucleus. Various seed storage proteins are cleaved once they have been transported into protein storage vacuoles; cleavage will bring on a new conformation necessary for their deposition (P09802).

In addition to these specific cleavages, more than 300 distinct types of amino-acid modifications have been identified to date in prokaryotic and eukaryotic proteins [80], of which each is specific to one or more amino acids. According to the RESID database (release 41) [80], cysteine is the amino acid which undergoes the most possible kinds of alterations (Fig. 3). One or more distinct types of modification can occur in a protein – and in various combinations – thus effectively extending the structural variety of a gene product. Many sequence modifications are irreversible, thus changing the property of the protein irreversibly too. Reversible protein modifications such as phosphorylation, S-nitrosylation or O-glycosylation can alter the dynamics of a protein and are considered to be major mechanisms for the fast and intricate regulation of metabolic enzymes [81] and hence signalling pathways [81–84]. Competition for different PTMs at a single site in response to distinct upstream signals, can provide a fine-tuning mechanism for signal integration. A good illustration for this kind of mechanism is the ‘histone modification code’, where the acetylation of Lys-9 in histone H3 would lead to transcriptional activation, or its tri-methylation would lead to transcriptional repression [85]. Transcription factors are also regulated through different and exclusive lysine-directed PTMs: Acetylation can stimulate or inhibit their activity, monoubiquitination generally enhances it, SUMOylation controls their subnuclear localization, and polyubiquitination signals their destruction by the proteasome [86].

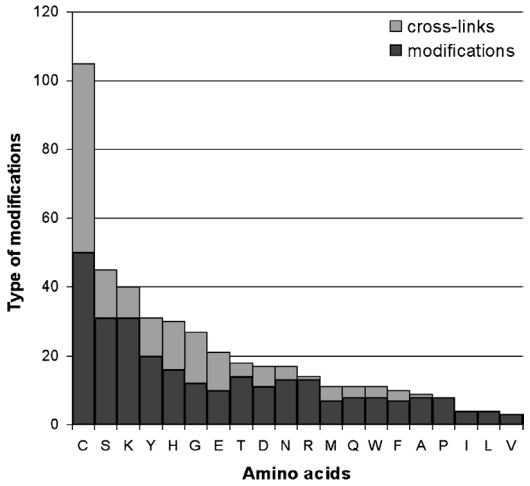

Fig. 3. Number of different types of amino-acid modifications and cross-links between distinct amino acids based on the RESID database (release 41).

Protein post-translational modification is a powerful mechanism to enhance the diversity of protein structures and to modify protein properties. Thus, it is hardly surprising that missing or abnormal modifications of a given protein can be the cause for dysfunction (see also Section 7) [87–90]. In particular, the highly dynamic modifications in glycoproteins are known to be associated with or are the reason for many distinct diseases [91]. Similarly, various abnormal post-translational modifications have been observed in aging tissue [92].

In Swiss-Prot, modified amino acids are annotated in the feature table. The type of feature key used depends on the nature of the modification and the exact name of the modified amino acid is indicated in the description field (Table 5). A list of the controlled vocabularies is available at http://www.expasy.org/sprot/userman.html#PTM_vocabularies. Modifications can also be described in the comment (CC) lines under the topic ‘PTM’. More than 40 keywords (KW) are used to

Table 5
Swiss-Prot annotation relevant to function-related protein modifications. Position-specific information is indicated in the feature table. Additional details can be provided in the comment (CC) line topic 'PTM'

P21591: Amidation of a methionine

```
KW   Amidation.
FT   MOD_RES      58     58       Methionine amide (G-59
FT                                provides amide group).
```

P19785: O-linked glycosylation

```
KW   Glycoprotein.
FT   CARBOHYD     10     10       O-linked (GlcNAc).
```

P07550: Palmitoylation of a cysteine

```
KW   Lipoprotein; Palmitate.
FT   LIPID        341    341      S-palmitoyl cysteine.
```

P09931: Disulfide bond linking 2 polypeptides

```
FT   DISULFID     29     29       Interchain (with C-8 in
FT                                small chain).
```

P14046: Thioester bond within a polypeptide

```
KW   Thioester bond.
FT   CROSSLNK     975    975      Isoglutamyl cysteine
FT                                thioester (Cys-Gln).
```

P11056: Binding site of a metal ligand

```
KW   Metal-binding; Iron.
FT   METAL         52     52      Iron (heme axial ligand).
```

P18485: Binding site of pyridoxal phosphate.

```
KW   Pyridoxal phosphate.
FT   BINDING      278    278      Pyridoxal phosphate
FT                                (covalent).
```

P46531: Supplementary free text annotation

```
CC   -!- PTM: Synthesized in the endoplasmic reticulum as an
CC       inactive form, which is proteolytically cleaved by a
CC       furin-like convertase in the trans-Golgi network
CC       before it reaches the plasma membrane to yield an
CC       active, ligand-accessible form. Cleavage results in a
CC       C-terminal fragment N(TM) and a N-terminal fragment
CC       N(EC). Following ligand binding, it is cleaved by
CC       TNF-alpha converting enzyme (TACE) to yield a membrane-
CC       associated intermediate fragment called notch
CC       extracellular truncation (NEXT). This fragment is then
CC       cleaved by presenilin dependent gamma-secretase to
CC       release a notch-derived peptide containing the
CC       intracellular domain (NICD) from the membrane (By
CC       similarity).
CC   -!- PTM: Phosphorylated (By similarity).
KW   Phosphorylation;
FT   SITE        1665   1666      Cleavage (by furin-like
FT                                protease) (By similarity).
```

describe protein modifications (see http://www.expasy.org/cgi-bin/get-entries?cat=PTM).

6. Protein–protein interactions

Cellular processes are generally carried out by protein assemblies rather than by individual proteins [93]. Protein complexes consist of at least two subunits, but can also be formed by dozens of polypeptides or more. Complex associations may last only a fraction of a millisecond but they can also form stable cellular structures. Many proteins are part of various complexes, each of which may act in a distinct functional context. As an example, the regulatory subunit of the cAMP-dependent protein kinase A (P10644) also interacts with the 40 kDa subunit of the replication factor C (P35250) [94]. A function is also frequently attributed to a complex rather than to the individual chains that make up the complex. In the case of homologous quaternary structures, the number of subunits can vary between taxonomic groups – as has been shown for DNA clamps which consist of a three-domain dimer in bacteria but a two-domain trimer in eukaryotes and Archea [95].

The formation of complexes is based on direct protein–protein interactions (PPIs) between the subunits, mediated by electrostatic interactions, hydrogen bonds, van der Waals attraction and hydrophobic effects. The minimal contact surface appears to be around 800 Å^2 [96] and the average interaction surface is 1600 ± 400 Å^2 [97]. The interaction strength between two proteins is quantitatively characterized by the equilibrium constant K_d. K_d values in the mM range are considered as rather weak, while values in the nM range or below are strong. In a biological context, several weak interactions between complex subunits can still contribute to a highly stable complex. Typical examples for transient interactions are enzyme – substrate reactions and interactions in signalling cascades. Permanent, stable complexes can be purified and eventually structurally analyzed as an assembly, examples are the ribosome [98], the RNA polymerase II [99], the ARP2/3 complex [100] (Fig. 4) or the proteasome [103].

The 'classical' types of contact between proteins are domain–domain interactions in which two independently folded domains – usually complementary in shape and charge – form the interface between the proteins [104]. In general, a central region protected from the solvent contributes the most to the binding energy. Sequence mutations in this region have a large impact on the interaction. Much research in domain–domain interaction has been carried out on serine proteases and their inhibitors. Another common contact type is the domain–peptide interaction: a domain of one protein interacts with a small portion of a second protein, which is unstructured in the absence of its binding partner. Typical examples are the binding between the major histocompatibility complex and the

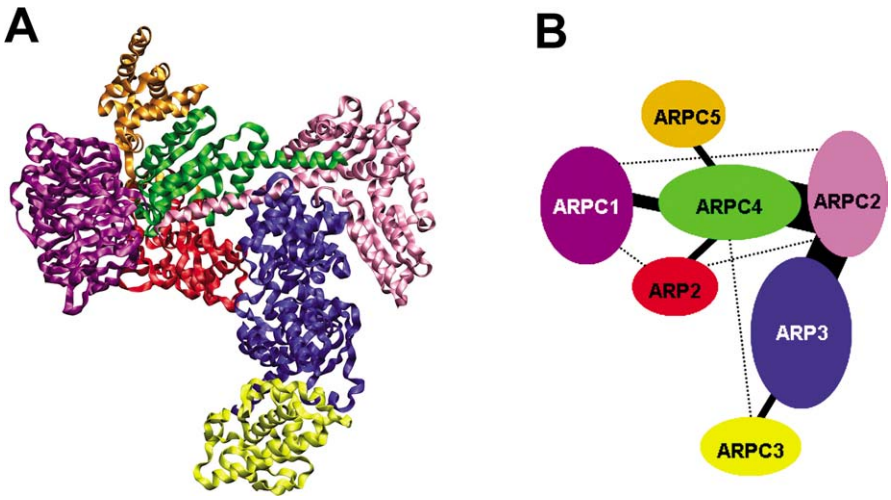

Fig. 4. The ARP2/3 complex. (**A**) Ribbon plot, generated from PDB 1K8K [100], shows the 7 subunits of the complex in distinct colours; (**B**) Schematic interaction diagram of the 7 subunits. Analysis of the complex structure reveals 6 binary interactions (threshold: $\geqslant$ 800 Å^2 contact surface), indicated with solid lines; the thickness of the lines corresponds to the buried contact surface area [96,100]. Genome-wide yeast 2-hybrid screens detected only one of these PPIs (IntAct, May 2005), see Table 7, example P33204, comment (CC) line topic 'Interaction'. Experimentally detected direct interactions [101,102], for which no contact surface can be observed in the structure, are indicated by dashed lines.

antigen, and regulatory modules of intracellular signalling cascades, like the SH2 domain which interacts with phosphotyrosine-containing target peptides in a sequence-specific manner. Protein–protein interactions can also induce conformational changes, as is often observed in signal transduction mechanisms. An example would be the tyrosine kinase Src, which contains SH2 and SH3 domains that mask the catalytic site of the enzyme. The Src-activating ligands contain specific SH2- and SH3-binding motifs; while the kinase and its ligand interact, the inhibition of the catalytic site is relieved [105].

For a long time, PPIs were studied individually using genetic, biochemical and biophysical techniques. More recently, genome-scale studies provide vast interaction datasets from high-throughput experiments. Yeast two-hybrid screens detect binary PPIs, without giving further information on functional complex associations [106–110]. Comparative analysis revealed not only an important amount of false-negative interactions (Fig. 4B), but led to the estimation that 30–50% of the reported interactions from large-scale projects do not exist [111–115]. Affinity purification with mass spectrometry detects the components of a complex without registering direct PPIs [116,117], however. Remarkably, stable complexes have rarely been found, and the composition and functionality of many of the detected transient complexes are not yet well understood. Proteome-wide interaction networks have been established for yeast [118], *C. elegans* [119] and fruit fly [109] and strive to assign functions to uncharacterized proteins. The yeast proteome consists of approximately 6200 proteins and would account for a minimum of 30 000 interactions. The number of existing PPIs is probably much higher since interactions occurring during distinct developmental stages or responses to different external conditions have also to be taken into account [111]. X-ray crystallography still provides the 'golden standard' in terms of accuracy for the structural analysis of protein complexes (Fig. 4B) but is not scalable to unravel a complete cell 'interactome'. Understanding the functional protein assembly in a cell will require an integrated approach by combining different experimental and theoretical methods [120].

The quaternary structure and any type of interaction with other proteins or protein complexes are described in Swiss-Prot entries. Cross-references to the IntAct database [121] facilitate access to complementary information including experimental details (Table 6).

Table 6
Swiss-Prot annotation relevant to PPIs. Literature-derived protein–protein interactions are annotated in the CC line topic 'SUBUNIT'. Binary interactions in the CC line topic 'INTERACTION' are automatically derived from the IntAct database. The FT keys 'REGION' and 'VARIANT' are used to specify an interaction

P33204: Stable complex association

```
CC   -!- FUNCTION: Part of the ARP2/3 complex implicated in
CC       the control of actin polymerization in cells (By
CC       similarity).
CC   -!- SUBUNIT: Belongs to the ARP2/3 complex composed of
CC       ARP2, ARP3, P41-ARC, P34-ARC, P21-ARC, P20-ARC and
CC       P16-ARC.(By similarity).
CC   -!- INTERACTION:
CC       P40518:ARC15; NbExp=1; IntAct=EBI-2757, EBI-2750;
DR   IntAct; P33204; -.
```

P40337: Complex formation prevented by a mutation

```
CC   -!- SUBUNIT: Part of a complex with elongin BC complex
CC       and hydroxylated HIF1A. Interacts with CUL2. Part of
CC       multisubunit CBC(VHL) E3 ubiquitin ligase complexes
CC       with elongin BC complex, CUL2 or CUL5 and RBX1.
CC       Interacts with chaperone CCT/TRiC. The interaction
CC       with CCT is required for the interaction with the
CC       elongin BC complex. Part of a complex with CCT and
CC       members of the Hsp70 chaperone family. Interacts with
CC       HIF1AN. Part of a complex with HIF1A, HIF1AN and
CC       HDAC1 or HDAC2 or HDAC3. The C-terminus interacts
CC       with CVBP1. Interacts with UBP33.
DR   IntAct; P40337; -.
FT   REGION      100    155       Involved in binding to CCT
FT                                complex.
FT   REGION      157    166       Interaction with elongin BC
FT                                complex.
FT   VARIANT     158    158       L -> P (in VHLD; type I-II;
FT                                abolishes release from
FT                                chaperonin complex and the
FT                                interaction with elongin BC
FT                                complex).
FT                                /FTId=VAR_005748.
```

7. Protein function and disease association

Once a polypeptide is functional, it participates in the coordination or maintenance of cellular processes, such as metabolism, transport, communication, growth, cellular biosynthesis or apoptosis. Dysfunction or, on the other hand, the simple lack of a given protein can lead to a disease status, but in many cases a combination of causes result in a disorder: genetic predisposition, environmental factors, infections and aging. Due to the complexity of living systems, understanding disease mechanisms may require moving from the global view of the organism as a whole to a more focused view of specific components at the molecular level.

Genetic mutations are one of the causes of protein alterations. Missense mutations, which ultimately lead to amino-acid substitutions, are the most frequent type of mutations related to disease [122,123]. A major challenge in medical genetics is to distinguish disease-causing missense mutations from neutral polymorphisms with no clinical relevance. Several criteria can be used to assess the pathogenicity of a mutation: de novo appearance of the mutation, segregation of the mutation with the disease within pedigrees, absence of the mutation in control individuals, change of amino-acid polarity or size in the protein, change in a domain which is conserved between species and/or shared between proteins belonging to the same family. If the function of the protein is known, the effect of the mutation can be assessed by in vitro mutagenesis and functional assay. Generally, missense mutations affect amino-acid residues with a relevant functional and structural role. They may have a deleterious effect in that they can lead to protein mistargeting, unstable miss-folded proteins, alteration of normally post-translationally modified sites, disruption of catalytic sites, or disruption of protein complexes. A mutation in the sulfatase modifying factor 1 (SUMF1; Q8NBK3) which activates all sulfatases (e.g., P15848, P15289, P08842) by transforming a conserved cysteine to 3-oxoalanine (Table 7) [124] is an example of a disorder based on missing PTMs. The lack of such a modification leads to multisulfatase deficiency (MSD), a severe multisystemic disorder which combines all the effects observed for each single sulfatase defect [125,126].

The mutation-disease relationship is not trivial. One same mutation can cause different phenotypes in individuals depending on the genome and the environment. Disease mutations in the fibroblast growth factor receptor (P21802) are a cause of certain craniosynostoses. In particular, the Cys342Tyr mutation is associated with two different craniosynostoses: Crouzon syndrome and Pfeiffer syndrome [127]. On the other hand, mutations in more than one gene may be necessary for disease manifestation, as shown for the Bardet–Biedl syndrome [128]. In order to elucidate the complex relationship between genotype and phenotype, great interest is devoted to the identification and cataloguing of single nucleotide polymorphisms (SNPs) that are expected to facilitate large-scale association genetics studies.

Swiss-Prot stores information related to protein function and detailed information is given on an enzyme's activity (Table 7). Disorders associated with the dysfunction of a protein are also indicated. Swiss-

P15289: Arylsulfatase

```
CC   -!- FUNCTION: Hydrolyzes cerebroside sulfate.
CC   -!- CATALYTIC ACTIVITY: A cerebroside 3 sulfate + H(2)O =
CC       a cerebroside + sulfate.
CC   -!- COFACTOR: Binds 1 magnesium ion per subunit.
CC   -!- DISEASE: Defects in ARSA are a cause of metachromatic
CC       leukodystrophy (MLD) [MIM:250100]. MLD is a disease
CC       characterized by intralysosomal storage of
CC       cerebroside-3-sulfate, as well as in the myelin
CC       membranes. Whereas storage occurs in many cells, the
CC       disease almost exclusively affects oligodendrocytes.
CC       Patients suffer from a progressive demyelination,
CC       which causes a variety of neurological symptoms,
CC       including gait disturbances, ataxias, optical
CC       atrophy, dementia, seizures, and spastic
CC       tetraparesis. Several years after the onset of the
CC       disease, patients die in a decerebrated state. Three
CC       forms of the disease can be distinguished according
CC       to the age at onset: late-infantile, juvenile and
CC       adult.
CC   -!- DISEASE: Arylsulfatase A activity is defective in
CC       multiple sulfatase deficiency (MSD) [MIM:272200]. MSD
CC       is a disorder characterized by decreased activity of
CC       all known sulfatases. MSD is due to defects in SUMF1
CC       resulting in the lack of post-translational
CC       modification of a highly conserved cysteine into 3-
CC       oxoalanine. It combines features of individual
CC       sulfatase deficiencies such as metachromatic
CC       leukodystrophy, mucopolysaccharidosis,
CC       chondrodysplasia punctata, hydrocephalus, ichthyosis,
CC       neurologic deterioration and developmental delay.
DR   MIM; 607574; -.
DR   MIM; 250100; -.
DR   MIM; 272200; -.
KW   Disease mutation; Hydrolase; Lipid metabolism; Magnesium;
KW   Metachromatic leukodystrophy; Metal-binding;
KW   Polymorphism; Sphingolipid metabolism.
FT   ACT_SITE    125    125
FT   ACT_SITE    125    125
FT   METAL         29     29       Magnesium.
FT   METAL         30     30       Magnesium.
FT   METAL         69     69       Magnesium (via 3-oxoalanine).
FT   METAL        281    281       Magnesium.
FT   METAL        282    282       Magnesium.
FT   MOD_RES       69     69       3-oxoalanine (Cys).
FT   VARIANT       76     76       L -> P.
FT                                 /FTId=VAR_007243.
FT   VARIANT       82     82       P -> L (in MLD; late-infantile onset;
FT                                 dbSNP:6151411).
FT                                 /FTId=VAR_007244.
FT   VARIANT       84     84       R -> Q (in MLD; mild).
FT                                 /FTId=VAR_007245.
FT   VARIANT       86     86       G -> D (in MLD; severe).
```

Prot aims to record all known single amino-acid substitutions, with an emphasis on human disease-related variants and their functional effects.

8. Concluding remarks

Even though Swiss-Prot entries generally correspond to a gene rather than a protein, it stores a wealth of information regarding protein variety and functional diversity. As shown for a number of proteins above, protein variants can be essential for an organism and protein dysfunctions can cause severe disorders. Information, such as sequence variety, protein complexes or PTMs that are not related to a conserved domain, is difficult to predict in a reliable fashion, – or indeed difficult to predict at all – and thus has to be annotated manually on the basis of experimental results. As knowledge on protein variety is essential for the understanding of cellular systems, Swiss-Prot aims to record all such data. For human entries, annotation on the naturally occurring polymorphisms and disease mutations provides valuable information on a protein's function. In Fig. 5, statistics on some relevant annotation is given based on the Swiss-Prot data from UniProt release 5.0 of 10 May 2005.

Table 7
Swiss-Prot annotation related to protein function and associated disorders. In the given example, the dysfunction of the SUMF1 impedes the activation of all sulfatases, thus causing MSD. In the Swiss-Prot entries, the protein function and related disease information is indicated in the comment (CC) lines. The sequence positions of variants are recorded in the feature table (FT). For disease mutations, the corresponding disorder is indicated in the description field of the feature key 'VARIANT'; if no additional information is given, the substitution refers to polymorphism. The keyword 'Polymorphism' is assigned to an entry if protein sequence variants are not associated with a disease status: the keyword 'Disease mutation' indicates disease-linked mutations. A specific disease term is used as keyword only when a disorder is associated with mutations in more than a single protein. A list of medical-oriented keywords is available at http://www.expasy.org/cgi-bin/get-entries?cat=Disease. Cross-references are provided to databases relevant to medical genetics such as Online Mendelian Inheritance in Men (OMIM) [129], dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), GenAtlas (http://www.genatlas.org/) and locus-specific databases

Q8NBK3: Sulfatase modifying factor 1 (SUMF1)

```
CC   -!- FUNCTION: Converts newly synthesized inactive
CC       sulfatases to their active form by modifying an
CC       active site cysteine residue to 3-oxoalanine. Known
CC       substrates include GALNS, ARSA, STS and ARSE.
CC   -!- DISEASE: Defects in SUMF1 are the cause of multiple
CC       sulfatase deficiency (MSD) [MIM:272200]. MSD is a
CC       clinically and
CC       biochemically heterogeneous disorder caused by the
CC       simultaneous impairment of all sulfatases, due to
CC       defective post-translational modification and
CC       activation. It combines features of individual
CC       sulfatase deficiencies such as metachromatic
CC       leukodystrophy, mucopolysaccharidosis,
CC       chondrodysplasia punctata, hydrocephalus, ichthyosis,
CC       neurologic deterioration and developmental delay.
CC       Inheritance is autosomal recessive.
DR   MIM; 607939; -.
DR   MIM; 272200; -.
KW   Disease mutation; Metachromatic leukodystrophy;
KW   Mucopolysaccharidosis; Polymorphism.
FT   VARIANT       20     20       L -> F (in MSD; loss of
FT                                 activity).
FT                                 /FTId=VAR_019050.
FT   VARIANT       63     63       S -> N.
FT                                 /FTId=VAR_016052.
FT   VARIANT       63     63       S -> N (in dbSNP:2819590).
FT                                 /FTId=VAR_016052.
FT   VARIANT      266    266       P -> L (in MSD; retains some
FT                                 activity).
FT                                 /FTId=VAR_019054.
```

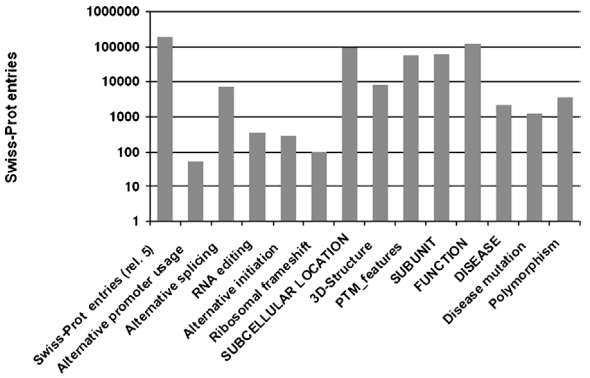

Fig. 5. Statistics on selected annotation items. The first column shows the number of entries in Swiss-Prot of UniProtKB release 5.0, other values on the x-axis indicate certain annotation items: terms in uppercase are comment (CC) line topics, terms in mixed case indicate keywords, except for 'PTM_features', which summarizes all feature keys relevant to post-translational amino-acid modifications.

Acknowledgements

The authors would like to thank Vivienne Baillie Gerritsen for the correction of the manuscript. We are grateful to Anne Estreicher for stimulating discussions on protein synthesis. This work is partially funded by the Swiss Federal Government through the State Secretariat for Education and Research and the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Apologies to authors whose work was not cited due to space limitation.

References

[1] A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, Swiss-Prot: juggling between evolution and stability, Brief. Bioinform. 5 (2004) 39–55.

[2] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, The Universal Protein Resource (UniProt), Nucleic Acids Res. 33 (2005) D154–D159.

[3] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370.

[4] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, A. Bairoch, ExPASy: The proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res. 31 (2003) 3784–3788.

[5] G.S. Shadel, D.A. Clayton, Mitochondrial DNA maintenance in vertebrates, Annu. Rev. Biochem. 66 (1997) 409–435.

[6] A.J. Bendich, Circular chloroplast chromosomes: the grand illusion, Plant Cell 16 (2004) 1661–1666.

[7] S. Baginsky, W. Gruissem, Chloroplast proteomics: potentials and challenges, J. Exp. Bot. 55 (2004) 1213–1220.

[8] A. Sakai, H. Takano, T. Kuroiwa, Organelle nuclei in higher plants: structure, composition, function, and evolution, Int. Rev. Cytol. 238 (2004) 59–118.

[9] D.B. Stern, M.R. Hanson, A. Barkan, Genetics and genomics of chloroplast biogenesis: maize as a model system, Trends Plant Sci. 9 (2004) 293–301.

[10] V.L. Stirewalt, C.B. Michalowski, W. Loeffelhardt, H.J. Bohnert, D.A. Bryant, Nucleotide sequence of the cyanelle DNA from *Cyanophora paradoxa*, Plant Mol. Biol. Rep. 13 (1995) 327–332.

[11] T.A. Ayoubi, W.J. Van De Ven, Regulation of gene expression by alternative promoters, FASEB J. 10 (1996) 453–460.

[12] N.J. Proudfoot, A. Furger, M.J. Dye, Integrating mRNA processing with transcription, Cell 108 (2002) 501–512.

[13] M. Deutsch, M. Long, Intron-exon structures of eukaryotic model organisms, Nucleic Acids Res. 27 (1999) 3219–3228.

[14] A.R. Kornblihtt, M. de la Mata, J.P. Federa, M.J. Munoz, G. Nogues, Multiple links between transcription and splicing, RNA 10 (2004) 1489–1498.

[15] R. Reed, Coupling transcription, splicing and mRNA export, Curr. Opin. Cell Biol. 15 (2003) 326–331.

[16] P.J. Lopez, B. Seraphin, YIDB: the Yeast intron database, Nucleic Acids Res. 28 (2000) 85–86.

[17] M.L. Bang, T. Centner, F. Fornoff, A.J. Geach, M. Gotthardt, M. McNabb, C.C. Witt, D. Labeit, C.C. Gregorio, H. Granzier, S. Labeit, The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system, Circ. Res. 89 (2001) 1065–1072.

[18] G. Ast, How did alternative splicing evolve?, Nat. Rev. Genet. 5 (2004) 773–782.

[19] D. Brett, H. Pospinil, J. Valcartel, J. Reich, P. Bork, Alternative splicing and genome complexity, Nat. Genet. 30 (2002) 29–30.

[20] S. Boue, I. Letunic, P. Bork, Alternative splicing and evolution, Bioessays 25 (2003) 1031–1034.

[21] L. Cartegni, S.L. Chew, A.R. Krainer, Listening to silence and understanding nonsense: exonic mutations that affect splicing, Nat. Rev. 3 (2002) 285–298.

[22] B. Modrek, A. Resch, C. Grasso, C. Lee, Genome-wide detection of alternative splicing in expressed sequences of human genes, Nucleic Acids Res. 29 (2001) 2850–2859.

[23] D. Schmucker, J.C. Clemens, H. Shu, C.A. Worby, J. Xiao, M. Muda, J.E. Dixon, S.L. Zipursky, Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity, Cell 101 (2000) 671–684.

[24] C. Caldas, C.W. So, A. MacGregor, A.M. Ford, B. McDonald, L.C. Chan, L.M. Wiedemann, Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene, Gene 208 (1998) 167–176.

[25] C. Finta, P.G. Zaphiropoulos, Intergenic mRNA molecules resulting from trans-splicing, J. Biol. Chem. 277 (2002) 5882–5890.

[26] S. Mass, A. Rich, Changing genetic information through RNA editing, Bioessays 22 (2000) 790–802.
[27] M. Schaub, W. Keller, RNA editing by adenosine deaminases generates RNA and protein diversity, Biochimie 84 (2002) 791–803.
[28] D.A. Campbell, S. Thomas, N.R. Sturm, Transcription in kinetoplastid protozoa: why be normal?, Microbes Infect. 5 (2003) 1231–1240.
[29] P. Vinciguerra, F. Stutz, mRNA export: an assembly line from genes to nuclear pores, Curr. Opin. Cell Biol. 16 (2004) 285–292.
[30] T.V. Pestova, I.B. Lomakin, J.H. Lee, S.K. Choi, T.E. Dever, C.U. Hellen, The joining of ribosomal subunits in eukaryotes requires eIF5B, Nature 403 (2000) 332–335.
[31] L.D. Kapp, J.R. Lorsch, The molecular mechanics of eukaryotic translation, Annu. Rev. Biochem. 73 (2004) 657–704.
[32] M. Kozak, Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6, EMBO J. 16 (1997) 2482–2492.
[33] N.N. Hashimoto, L.S. Carnevalli, B.A. Castilho, Translation initiation at non-AUG codons mediated by weakened association of eukaryotic initiation factor (eIF) 2 subunits, Biochem. J. 367 (2002) 359–368.
[34] S. Malarkannan, T. Horng, P.P. Shih, S. Schwab, N. Shastri, Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism, Immunity 10 (1999) 681–690.
[35] C. Touriol, S. Bornes, S. Bonnal, S. Audigier, H. Prats, A.C. Prats, S. Vagner, Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons, Biol. Cell. 95 (2003) 169–178.
[36] O. Namy, J.P. Rousset, S. Napthine, I. Brierley, Reprogrammed genetic decoding in cellular gene expression, Mol. Cell. 13 (2004) 157–168.
[37] P.V. Baranov, R.F. Gesteland, J.F. Atkins, Recoding: translational bifurcations in gene expression, Gene 286 (2002) 187–201.
[38] I.P. Ivanov, S. Matsufuji, Y. Murakami, R.F. Gesteland, J.F. Atkins, Conservation of polyamine regulation by translational frameshifting from yeast to mammals, EMBO J. 19 (2000) 1907–1917.
[39] K. Shigemoto, J. Brennan, E. Walls, C.J. Watson, D. Stott, P.W. Rigby, A.D. Reith, Identification and characterisation of a developmentally regulated mammalian gene that utilises −1 programmed ribosomal frameshifting, Nucleic Acids Res. 29 (2001) 4079–4088.
[40] D.L. Hatfield, V.N. Gladyshev, How selenium has altered our understanding of the genetic code, Mol. Cell. Biol. 22 (2002) 3565–3576.
[41] B. Cobucci-Ponzano, M. Rossi, M. Moracci, Recoding in archaea, Mol. Microbiol. 55 (2005) 339–348.
[42] P. Schimmel, K. Beebe, Molecular biology: genetic code seizes pyrrolysine, Nature 431 (2004) 257–258.
[43] R.A. Bradshaw, W.W. Brickey, K.W. Walker, N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families, Trends Biochem. Sci. 23 (1998) 263–267.
[44] B.B. Quimby, A.H. Corbett, Nuclear transport mechanisms, Cell. Mol. Life Sci. 58 (2001) 1766–1773.
[45] N. Wiedemann, A.E. Frazier, N. Pfanner, The protein import machinery of mitochondria, J. Biol. Chem. 279 (2004) 14473–14476.
[46] C.M. Koehler, New developments in mitochondrial assembly, Annu. Rev. Cell Dev. Biol. 20 (2004) 309–335.
[47] A. Chacinska, S. Pfannschmidt, N. Wiedemann, V. Kozjak, L.K. Sanjuan Szklarz, A. Schulze-Specking, K.N. Truscott, B. Guiard, C. Meisinger, N. Pfanner, Essential role of Mia40 in import and assembly of mitochondrial intermembrane space proteins, EMBO J. 23 (2004) 3735–3746.
[48] R.A. Stuart, Insertion of proteins into the inner membrane of mitochondria: the role of the Oxa1 complex, Biochim. Biophys. Acta 1592 (2002) 79–87.
[49] M. Preuss, M. Ott, S. Funes, J. Luirink, J.M. Herrmann, Evolution of mitochondrial oxa proteins from bacterial YidC, inherited and acquired functions of a conserved protein insertion machinery, J. Biol. Chem. 280 (2005) 13004–13011.
[50] A. Nada, J. Soll, Inner envelope protein 32 is imported into chloroplasts by a novel pathway, J. Cell Sci. 117 (2004) 3975–3982.
[51] J. Soll, E. Schleiff, Protein import into chloroplasts, Nat. Rev. Mol. Cell. Biol. 5 (2004) 198–208.
[52] D. Schuenemann, Structure and function of the chloroplast signal recognition particle, Curr. Genet. 44 (2004) 295–304.
[53] C. Robinson, A. Bolhuis, Tat-dependent protein targeting in prokaryotes and chloroplasts, Biochim. Biophys. Acta 1694 (2004) 135–147.
[54] R.J. Keenan, D.M. Freymann, R.M. Stroud, P. Walter, The signal recognition particle, Annu. Rev. Biochem. 70 (2001) 755–775.
[55] T.A. Rapoport, V. Goder, S.U. Heinrich, K.E.S. Matlack, Membrane-protein integration and the role of the translocation channel, Trends Cell Biol. 14 (2004) 568–575.
[56] B. Martoglio, B. Dobberstein, Signal sequences: more than just greasy peptides, Trends Cell Biol. 8 (1998) 410–415.
[57] W. Schliebs, W.-H. Kunau, Peroxisome membrane biogenesis: the stage is set, Curr. Biol. 14 (2004) R397–R399.
[58] S.J. Gould, C.S. Collins, Peroxisomal protein import: is it really that complex?, Nat. Rev. Mol. Cell Biol. 3 (2002) 382–389.
[59] P.B. Lazarow, Peroxisome biogenesis: advances and conundrums, Curr. Opin. Cell Biol. 15 (2003) 489–497.
[60] J. Imai, H. Yashiroda, M. Maruya, I. Yahara, K. Tanaka, Proteasomes and molecular chaperones: cellular machinery responsible for folding and destruction of unfolded proteins, Cell Cycle 2 (2003) 585–590.
[61] A. Komeili, E.K. O'Shea, New perspectives on nuclear transport, Annu. Rev. Genet. 35 (2001) 341–364.
[62] R. Schekman, L. Orci, Coat proteins and vesicle budding, Science 271 (1996) 1526–1533.
[63] J. Lippincott-Schwartz, T.H. Roberts, K. Hirschberg, Secretory protein trafficking and organelle dynamics in living cells, Annu. Rev. Cell Dev. Biol. 16 (2000) 557–589.
[64] M. Sinensky, Recent advances in the study of prenylated proteins, Biochim. Biophys. Acta 1529 (2000) 203–209.

[65] R.E. Dalbey, A. Kuhn, Evolutionarily related insertion pathways of bacterial, mitochondrial, and thylakoid membrane proteins, Annu. Rev. Cell Dev. Biol. 16 (2000) 51–87.
[66] C.R. Sanders, J.K. Myers, Disease-related misassembly of membrane proteins, Annu. Rev. Biophys. Biomol. Struct. 33 (2004) 25–51.
[67] T.R. Kau, J.C. Way, P.A. Silver, Nuclear transport and cancer: from mechanism to intervention, Nat. Rev. Cancer 4 (2004) 106–117.
[68] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[69] C.M. Dobson, Principles of protein folding, misfolding and aggregation, Semin. Cell & Dev. Biol. 15 (2004) 3–16.
[70] F.U. Hartl, J. Martin, Molecular chaperones in cellular protein folding, Curr. Opin. Struct. Biol. 5 (1995) 92–102.
[71] A. Sali, E. Shakhnovich, M. Karplus, How does a protein fold?, Nature 69 (1994) 248–251.
[72] Y. Zhou, M. Karplus, Interpreting the folding kinetics of helical proteins, Nature 401 (1999) 400–403.
[73] C.A. Orengo, F.M. Pearl, J.E. Bray, A.E. Todd, A.C. Martin, L. Lo Conte, J.M. Thornton, The CATH Database provides insights into protein structure/function relationships, Nucleic Acids Res. 27 (1999) 275–279.
[74] L. Lo Conte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, C. Chothia, SCOP: a structural classification of proteins database, Nucleic Acids Res. 28 (2000) 257–259.
[75] M.A. Andrade, C. Perez-Iratxeta, C.P. Ponting, Protein repeats: structures, functions and evolution, J. Struct. Biol. 134 (2001) 117–131.
[76] M. Shen, F.P. Davis, A. Sali, The optimal size of a globular protein domain: A simple sphere-packing model, Chem. Phys. Lett. 405 (2005) 224–228.
[77] R.R. Copley, I. Letunic, P. Bork, Genome and protein evolution in eukaryotes, Curr. Opin. Chem. Biol. 6 (2002) 39–45.
[78] L. Patthy, Genome evolution and the evolution of exon-shuffling – a review, Gene 238 (1999) 103–114.
[79] E.V. Kriventseva, I. Koch, R. Apweiler, M. Vingron, P. Bork, M.S. Gelfand, S. Sunyaev, Increase of functional diversity by alternative splicing, Trends Genet. 19 (2003) 124–128.
[80] J.S. Garavelli, The RESID Database of Protein Modifications as a resource and annotation tool, Proteomics 4 (2004) 1527–1533.
[81] S.C. Huber, S.C. Hardin, Numerous posttranslational modifications provide opportunities for the intricate regulation of metabolic enzymes at multiple levels, Curr. Opin. Plant Biol. 7 (2004) 318–322.
[82] J. Seo, K.J. Lee, Post-translational modifications and their biological functions: proteomic analysis and systematic approaches, J. Biochem. Mol. Biol. 37 (2004) 35–44.
[83] C. Brahimi-Horn, N. Mazure, J. Pouyssegur, Signalling via the hypoxia-inducible factor-1alpha requires multiple post-translational modifications, Cell Signal 17 (2005) 1–9.
[84] T.L. Tootle, I. Rebay, Post-translational modifications influence transcription factor activity: a view from the ETS superfamily, Bioessays 27 (2005) 285–298.
[85] C.L. Peterson, M.A. Laniel, Histones and histone modifications, Curr. Biol. 14 (2004) R546–R551.
[86] R.N. Freiman, R. Tjian, Regulating the regulators: lysine modifications make their mark, Cell 112 (2003) 11–17.
[87] J.U. Baenziger, A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease, Cell 113 (2003) 421–422.
[88] C.X. Gong, F. Liu, I. Grundke-Iqbal, K. Iqbal, Post-translational modifications of tau protein in Alzheimer's disease, J. Neural. Transm. 112 (2005) 813–838.
[89] S.M. Anderton, Post-translational modifications of self antigens: implications for autoimmunity, Curr. Opin. Immunol. 16 (2004) 753–758.
[90] C. Schoneich, Mass spectrometry in aging research, Mass Spectrom. Rev., in press.
[91] T. Marquardt, J. Denecke, Congenital disorders of glycosylation: review of their molecular bases, clinical presentations and specific therapies, Eur. J. Pediatr. 162 (2003) 359–379.
[92] P.A. Cloos, S. Christgau, Post-translational modifications of proteins: implications for aging, antigen recognition, and autoimmunity, Biogerontology 5 (2004) 139–158.
[93] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Garland, New York, 2002.
[94] R.S. Gupte, Y. Weng, L. Liu, M.Y. Lee, The second subunit of the replication factor C complex (RFC40) and the regulatory subunit (RIalpha) of protein kinase A form a protein complex promoting cell survival, Cell. Cycle 4 (2005) 323–329.
[95] P. Aloy, R.B. Russell, The third dimension for protein interactions and complexes, Trends Biochem. Sci. 27 (2002) 633–638.
[96] A.M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, M. Gerstein, Bridging structural biology and genomics: assessing protein interaction data with known complexes, Trends Genet. 18 (2002) 529–536.
[97] S.J. Wodak, J. Janin, Structural basis of macromolecular recognition, Adv. Protein Chem. 61 (2002) 9–73.
[98] M.M. Yusupov, G.Z. Yusupova, A. Baucom, K. Lieberman, T.N. Earnest, J.H. Cate, H.F. Noller, Crystal structure of the ribosome at 5.5 A resolution, Science 292 (2001) 883–896.
[99] P. Cramer, D.A. Bushnell, R.D. Kornberg, Structural basis of transcription: RNA polymerase II at 2.8-angstrom resolution, Science 292 (2001) 1863–1876.
[100] R.C. Robinson, K. Turbedsky, D.A. Kaiser, J.B. Marchand, H.N. Higgs, S. Choe, T.D. Pollard, Crystal structure of Arp2/3 complex, Science 294 (2001) 1679–1684.
[101] L.M. Machesky, R.H. Insall, Scar1 and the related Wiskott-Aldrich syndrome protein, WASP, regulate the actin cytoskeleton through the Arp2/3 complex, Curr. Biol. 8 (1988) 1347–1356.
[102] R.D. Mullins, W.F. Stafford, T.D. Pollard, Structure, subunit topology, and actin-binding activity of the Arp2/3 complex from Acanthamoeba, J. Cell Biol. 136 (1997) 331–343.
[103] M. Unno, T. Mizushima, Y. Morimoto, Y. Tomisugi, K. Tanaka, N. Yasuoka, T. Tsukihara, The structure of the mammalian 20S proteasome at 2.75-Å resolution, Structure 10 (2002) 609–618.
[104] R.C. Liddington, Structural basis of protein–protein interactions, Methods Mol. Biol. 261 (2004) 3–14.

[105] R. Roskoski Jr., Src protein-tyrosine kinase structure and regulation, Biochem. Biophys. Res. Commun. 324 (2004) 1155–1164.

[106] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, Nature 403 (2000) 623–627.

[107] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl Acad. Sci. USA 98 (2001) 4569–4574.

[108] J.C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, P. Legrain, The protein–protein interaction map of *Helicobacter pylori*, Nature 409 (2001) 211–215.

[109] L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley Jr., K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shimkets, M.P. McKenna, J. Chant, J.M. Rothberg, A protein interaction map of *Drosophila melanogaster*, Science 302 (2003) 1727–1736.

[110] S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.W. Cusick, F.P. Roth, D.E. Hill, M. Vidal, A map of the interactome network of the metazoan *C. elegans*, Science 303 (2004) 540–543.

[111] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399–403.

[112] P. Aloy, R.B. Russell, Potential artefacts in protein-interaction networks, FEBS Lett. 530 (2002) 253–254.

[113] C.M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, Protein interactions: two methods for assessment of the reliability of high throughput observations, Mol. Cell Proteomics 1 (2002) 349–356.

[114] G.D. Bader, C.W. Hogue, Analyzing yeast protein–protein interaction data obtained from different sources, Nat. Biotechnol. 20 (2002) 991–997.

[115] E. Sprinzak, S. Sattath, H. Margalit, How reliable are experimental protein–protein interaction data?, J. Mol. Biol. 327 (2003) 919–923.

[116] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, M. Tyers, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, Nature 415 (2002) 180–183.

[117] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (2002) 141–147.

[118] B. Schwikowski, P. Uetz, S. Fields, A network of protein–protein interactions in yeast, Nat. Biotechnol. 18 (2000) 1257–1261.

[119] A.J. Walhout, S.J. Boulton, M. Vidal, Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm, Yeast 17 (2000) 88–94.

[120] A. Sali, R. Glaeser, T. Earnest, W. Baumeister, From words to literature in structural proteomics, Nature 422 (2003) 216–225.

[121] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, IntAct: an open source molecular interaction database, Nucleic Acids Res. 32 (2004) D452–D455.

[122] The Human Gene Mutation Database, http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html.

[123] S.E. Antonarakis, D.N. Cooper, Mutations in human genetic disease, in: D.N. Cooper (Ed.), Encyclopedia of the Human Genome, Nature Publishing Group, London, 2003, pp. 227–253.

[124] B. Schmidt, T. Selmer, A. Ingendoh, K. von Figura, A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency, Cell 28 (1995) 271–278.

[125] T. Dierks, B. Schmidt, L.V. Borissenko, J. Peng, A. Preusser, M. Mariappan, K. von Figura, Multiple sulfatase deficiency is caused by mutations in the gene encoding the human C(alpha)-formylglycine generating enzyme, Cell 113 (2003) 435–444.

[126] M.P. Cosma, S. Pepe, I. Annunziata, R.F. Newbold, M. Grompe, G. Parenti, A. Ballabio, The multiple sulfatase deficiency gene encodes an essential and limiting factor for the activity of sulfatases, Cell 113 (2003) 445–456.

[127] P. Rutland, L.J. Pulleyn, W. Reardon, M. Baraitser, R. Hayward, B. Jones, S. Malcolm, R.M. Winter, M. Oldridge, S.F. Slaney, M.D. Poole, A.O.M. Wilkie, Identical mutations in the FGFR2 gene cause both Pfeiffer and Crouzon syndrome phenotypes, Nat. Genet. 9 (1995) 173–176.

[128] N. Katsanis, S.J. Ansley, J.L. Badano, E.R. Eichers, R.A. Lewis, B.E. Hoskins, P.J. Scambler, W.S. Davidson, P.L. Beales, J.R. Lupski, Triallelic inheritance in Bardet–Biedl syndrome, a Mendelian recessive disorder, Science 293 (2001) 2256–2259.

[129] Online Mendelian Inheritance in Man, OMIM (TM), McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000, http://www.ncbi.nlm.nih.gov/omim/.