FEBS Letters 498 (2001) 214–218 FEBS 24915

Minireview

A genomic perspective on human proteases

Christopher Southan

Department of Computational Biology, Gemini Genomics (UK) Ltd, 162 Science Park, Milton Road, Cambridge CB4 0GH, UK

Received 10 April 2001; accepted 27 April 2001

First published online 16 May 2001

Abstract Over 400 human proteases documented in secondary databases can already be delineated in genomic sequence. A Genome Ontology annotation of 30 585 sequences in the provisional human proteome set recognises 498 proteases, i.e. 1.6%. Homology searches against finished sequence and comparisons between mouse and zebrafish are likely to increase this total. However, the data already indicate that the mechanistic class, sequence family and domain distribution of the genomic complement of proteases is unlikely to shift significantly from that already observed in the transcript data. Genomically derived novel sequences will require bioinformatic analysis and biochemical verification. The increasing availability of annotated genomic data will enable studies of splice variants, transcriptional control, polymorphisms, pseudogenes, inactive homologues and evolution. Comparative work on complete human protease families should produce a more integrated picture of their biochemistry and physiology. Genomic data will also lead to the identification of new protease involvement in disease processes and their evaluation as drug targets.

Key words: Protease; Exon; Gene organisation; Annotation; Domain

1. Introduction

Human genome data will impact protease research in five main areas. The first is the ability to study, in silico, in vitro and in vivo all paralogous members of a human protease family. The study of individual proteases often leaves questions as to the extent of unique or partitioned biochemical roles as opposed to functional redundancy. Examples of parallel studies include the expression analysis of all 15 human kallikrein sequences on 19q13 [1], comparisons between BACE on 11q22.2 and BACE2 on 21q22.3 [2,3], or the two angiotensin-converting enzymes, ACE1 from 17q23 and ACE2 from Xp22 [4]. With the caveat of the small remaining gaps, the genomic complement of these three families is likely to be complete. The second impact will be the opportunity to study species orthologues, not only for the similarities and differences in physiological context between human, mouse and fish but also, in some cases, conserved biochemical function compared to ancestral homologues in yeast, worm or fly. The same examples can be used here; mice have a larger number of kallikrein genes, at least some of which may not have true human orthologues [1]. BACE1 and BACE2 have only single rat and mouse orthologues but share at least one common ancestor in fish [5]. Similarly ACE1 and ACE2 each have rodent orthologues but ACE-like homologues can be traced back to the fly [6].

The third impact will be in the study of gene structure, such as exon/intron organisation, splice variants, transcriptional regulation, mutations, polymorphisms and evolutionary history. Again, these can include comparisons between human, mouse and fish. The fourth area will be investigating the roles of the ancillary functional sequence units, such as signal peptides, propeptides, transmembrane helices, lipid anchor regions or interaction domains that are found in association with catalytic modules. The multiplicity and permutations of these are a defining charateristic of vertebrate proteases. Last but not least, the availability of the predicted proteome will lead to the identification of new biochemically important protease substrates and/or endogenous modulators of their activity.

2. Genomic and transcript data sources

Monitoring the avalanche of data entering the GenBank/EMBL divisions for new protease sequences is a daunting task. However, the biological community is increasingly well provided with secondary databases that extract and curate subsets from the primary data. Many of these have utility for the study of proteases but only a few personal favourites can be described here. Top of this list is the MEROPS protease database that provides a regularly updated compendium of all proteases extracted from the primary databases [7]. As well as grouping these at the species and sequence family level the database also provides the internationally accepted mechanism and alignment-based protease classification system that will be used in this article. The RefSeq resource maintains a non-redundant set of human mRNAs, which are mapped onto genomic sequence via LocusLink [8]. The SwissProt/TrEMBL (SP-TR) protein database includes a non-redundant set of human sequences [9]. InterPro is a comprehensive protein family and domain database, which provides automated annotation for all of SP-TR including the human proteome set [10]. Ensembl is a major project for providing homology-supported predicted gene products from a 'golden path' of assembled human genome data [11]. This is regularly updated for viewing gene models and many other features extracted from the sequence data, including InterPro matches for the predicted proteins.

So how do these data sources relate to each other for assessing proteases in a genomic context? Most human entries

E-mail: chris.southan@gemini-genomics.com

PII: S0014-5793(01)02490-5

in MEROPS can be linked to a SP-TR protein sequence and a RefSeq mRNA. The protease classifications in MEROPS largely correspond to those derived by InterPro. However, some family divisions made on the basis of manually aligned catalytic regions in MEROPS, may be merged as a consequence of the use of automated scoring thresholds against the domain databases underpinning InterPro. The unique advantage of the latter is the identification and graphical depiction of all the known domains, in addition to the catalytic modules (see Fig. 1) [10,12]. For visualising genomic context, most of the MEROPS or InterPro entries can now be fine-mapped (i.e. unequivocally aligned at the exon level) onto genomic sequence, via either LocusLink or Ensembl. The latter includes those proteases not represented as mRNA entries in RefSeq but have enough similarity matches to known sequences to support a predicted gene model. An example of the progress already made in genome annotation is given in Fig. 1.

3. Genomic and transcript protease numbers (as of March, 2001)

Although the gene product inventory is some way from completion, the recent human genome paper included the first compilation of an integrated protein index of 31 778 protein sequences [11]. An updated set of 30 585 sequences was released on the European Bioinformatics Institute website. This includes 15 691 SP-TR proteins and an additional 14 894 proteins predicted by Ensembl. The entire set was analysed by InterPro for domain and protein family composition. The non-redundant mRNA total in RefSeq now stands at 15 752. This means that approximately half of all human proteins are already represented as full-length coding sequences.

Defining the number of proteases in silico, from transcript or genomic data, as well as detecting their possible evolutionary relationships, requires the application of scoring thresholds to the results of searching or alignment algorithms. Although the practical choice of these methods lies outside the scope of this review, it is important to note that the secondary databases interrogated for this article, MEROPS, InterPro and Ensembl, assign both domain and family homology conservatively. For such large-scale operations there are many good reasons to stay out of the 'twilight zone'. One advantage for proteases, is that the inferred biochemical properties, automatically or manually transferred during annotation of a new sequence, should be reliably linked to the small (and shrinking in relative numbers) subset of proteins whose catalytic activity has been experimentally demonstrated. In expert hands, the targeted use of low-similarity thresholds can certainly extend the homology identification envelope to include, for example, a newly extended superfamily of predicted cysteine proteases that includes five *Caenorhabditis elegans* and four human paralogues [13].

In terms of absolute numbers, the MEROPS database currently lists 405 human sequences. This corresponds to 2.9% of known human proteins but only 1.3% of the 30 585 genome set. A preliminary Genome Ontology annotation of this set classified 498 proteases i.e. 1.6% (see www.ebi.ac.uk/proteome/). The InterPro annotations of the Ensembl release include 297 proteases in the 25 790 confirmed gene set i.e. about 1.2%. These relative numbers are significantly below the 1.8% estimate made recently which was largely based on model organism data [14]. However, there are technical reasons why these initial genome-wide automated annotations may underestimate the protease total. Between now and the completion of the human data, the increase in mammalian mRNA and EST coverage, together with exon comparisons between human, mouse and zebrafish, are likely to increase the protease total. Given that the probability of a human sequence being cloned is, to very rough approximation, proportional to its mRNA abundance, it can be speculated that those new genome-derived proteases confirmed as real transcripts may be low-abundance or and/or tissue-specific gene products.

4. Mechanism, sequence family, and domain distribution

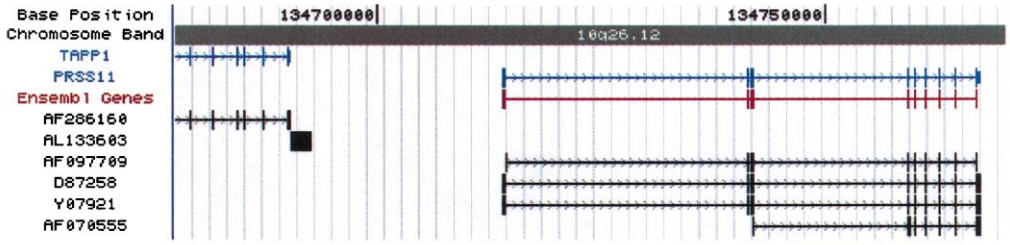
Although the human sequences in MEROPS approximately doubled between 1998 and 2000, there was relatively little shift in the mechanistic class distribution. These currently stand at 3% aspartic, 23% cysteine, 36% metallo, and 32% serine. The four largest families, the S1 trypsin–serine proteases, the M12 ADAM metalloproteases, the C12/C19 ubiquitin-specific proteases and M10 matrix metalloproteases, have undergone major expansions over this period. These transcript-derived totals stand at 101, 35, 31, and 24 members respectively. However, comparison with InterPro annotation of the proteome set, including for example 118 trypsin proteases, indicates that the expansion of these families by contributions from new genomic predictions, is modest. This suggests that the final protease set will have a broadly similar family composition to those already represented in transcript data. The human genome paper also reported that the trypsin-like (S1) serine proteases are associated with 18 other domains [11]. Although none of these were novel, it seems likely that expert analysis of the type recently reported for the protease-associated domain may reveal new combinations [15]. For those domains that are annotated the InterPro query tools enable interrogation of their species and sequence family combinations [10,12].

5. Inactive homologues and pseudogenes

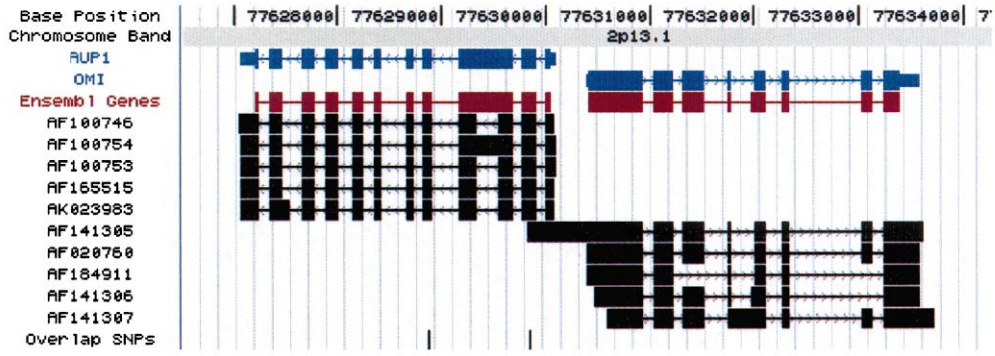
The MEROPS database has assigned 12% of the protease sequences as inactive homologues using the criteria of residue substitutions within critical active-site regions. However, it should be borne in mind that, from the vast number of protease homologues predicted from nucleic acid sequences, only a minority have been experimentally proven to be catalytically active. In fact only 18% of the 1634 protease families across all organisms currently in MEROPS have an Enzyme Commission number for one of their members. Nearly all of the major protease families include at least one inactive paralogue. The ADAM (M12) family has the largest proportional representation of 34%. Interestingly, these are at different chromosomal locations. Any new inactive homologues revealed in genomic data are worthy of biochemical investigation because they can have the same functional set of ancillary domains as their catalytically active paralogues.

Pseudogenes in genomic DNA result from reverse transcription from a mRNA transcript (processed pseudogenes) or from gene duplication and subsequent degradation (non-processed pseudogenes). Although these are considered 'dead', the RefSeq human mRNA entries include over 1000 tran-

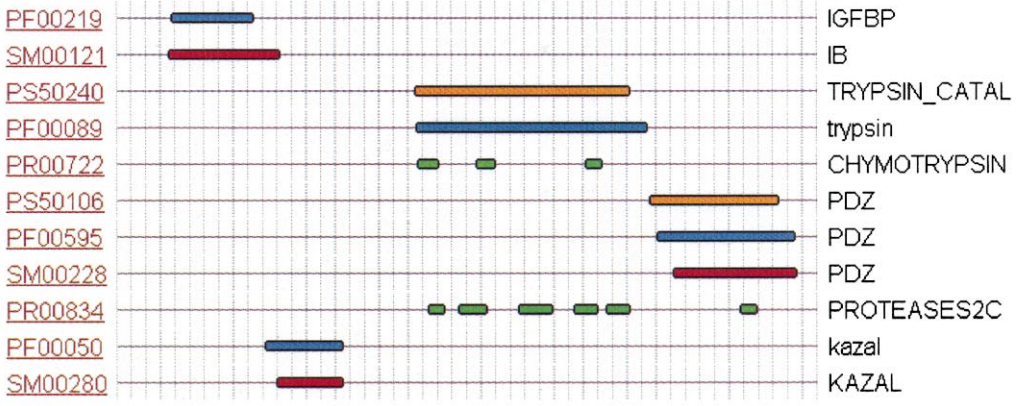
(a)



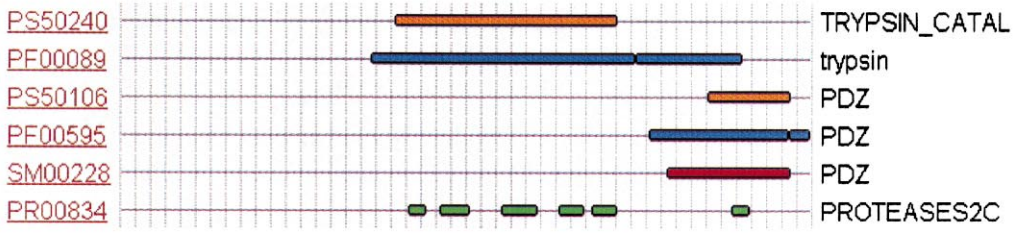
(b)



(c)



(d)



scribed pseudogenes. Examples of protease pseudogenes have recently been reported within the kallikrein and tryptase gene clusters on 19q13 and 16p13 respectively [1,16]. An example of an actively transcribed protease pseudogene is the (pro) napsin-B gene [17]. The mRNA is found exclusively in cells related to the immune system but lacks an in-frame stop codon and contains a number of polymorphisms, one of which replaces a catalytically crucial Gly residue with an Arg. It will require the data quality level of finished genomic sequence before such subtle differences between translated and non-translated proteases can be discerned with confidence in silico.

Fig. 1. Genomic context and protein features for the HtrA1 and HtrA2 proteases. Panels (a) and (b) include selected genomic viewer features on the left-hand side (background information, including additional display options, can be accessed at genome.cse.ucsc.edu/ and www.ensembl.org/). For each of the genes in (a) and (b) the immediate (left-hand, irrespective of orientation) neighbouring gene is included. From top to bottom, the features are; the base position, which refers to the coordinates in the golden path assembly, used by both the University of Santa Cruz and Ensembl teams (the two panels are not to scale as (a) is a longer gene); the cytogenetic band is immediately underneath. The blue lines are gene structures as confirmed by the consensus of the alignments between genomic and mRNAs. The brown lines are the confirmed gene models from Ensembl. The black numbers are mRNA primary accession numbers that have exon matches to the genomic DNA, e.g. The HtrA1/PRSS11 has four mRNA entries and the HtrA2/Omi gene has five. Panel (b) shows two SNPs in exons of the AUP1 gene. Examination of the graphically depicted features for these genes and following the links back to database entries and publications reveals many subtleties but only a few can be highlighted here. HrA1 and 2 are paralogues, sharing 45% amino acid identity across 320 residues, but both their gene structures, orientations and neighbouring genes are different. HtrA1 shows two 5′ exons, separated by large introns that are lacking in HtrA2. Given that AF097709 is slightly shorter, the exons agree with D87256 and Y07921. AF070555, a partial sequence from infant brain, shows exon 4 extended by seven bases but this clone does not translate into a complete protein. The more compact gene structure of HrA2 (b) is in opposite orientation i.e. the longest exon (8) is on the 3′ end. The gene structure is supported by five mRNAs but four of these show alternative transcripts; AF18491 lacking exons 2 and 6, AF141306, loss of exon 2 plus an extended exon 4, and AF141306, the fusion between exons 4 and 5. Unusually, the 3′ end of AF143065 overlaps with the two 5′-most exons of the neighbouring AUP1 gene. Panels (c) and (d) show the depiction of protein domains in HtrA1 and HtrA2. The graphical displays were retrieved from InterPro (www.ebi.ac.uk/interpro/). The domain databases used within InterPro, listed on the left side, are; PS = PROSITE, PF = Pfam, SM = SMART, PR = PRINTS. Both gene products are homologous to the C-terminal (3′ end) of bacterial heat-shock proteases, a relationship detected within InterPro by the combination of a trypsin and PDZ domain as well as a match to the proteases 2C family pattern in PRINTS (this has since been re-classified as family S1C in MEROPS) [12]. However, panel (c) shows that HtraA1 has two additional N-terminal features, the insulin growth factor-binding protein (IGFBP), and a Kazal protease inhibitor motif. These correspond to the three 5′ most exons in the gene. The particular four-domain combination in HrA1 is, so far, only detectable in mammals suggesting that the two N-terminal domains may have been 'shuffled-in' after duplication of a common ancestor. The combined utility of the annotated gene and protein features, in this example, is the ability to predict accurately which protein domains in (d) might be affected by the complex exon-splicing pattern observed in (b).
←

6. Alternative splicing and polyadenylation

Alternative transcript-splicing is predicted to affect at least 35% of all human gene products and the extent of alternative polyadenylation positions within the 3′ terminal exons remains unknown [18]. Both processes introduce additional layers of complexity into protease biology. For example, all the numbers quoted in Section 4, which were implicitly one gene–one mRNA–one protein, have to be expanded by 40% to give a more plausible picture of transcript permutations, without even considering differential polyadenylation. To use a long-standing example it is only recently that a physiological basis has been proposed to explain the expression of the enzymatically equivalent somatic and germinal isoforms of ACE [19]. Two cases have recently been described where tissue-specific removal of introns near active sites can give rise to catalytically inactive proteins but where ancillary interaction or anchoring domains may still be active. The first of these is a novel BACE mRNA lacking a 44-amino acid region from exon 3, located between the two catalytic aspartyl residues [20]. This is expressed as a pancreas-specific splice variant but is absent in brain. These findings can explain the previously observed paradox of high BACE transcription in pancreas with very low enzymatic activity but raises the question of what functional role this splice variant might have. To add to the transcript complexity of BACE there are at least three forms observed in multiple tissue Northern blots and public mRNA sequences that arise from different polyadenlylation positions in the 3′ UTR rather than alternative exon usage [21].

The second case is HtrA2, a protein related to the bacterial heat-shock proteases, consisting of eight exons on 2p13.1 [22]. There are two reports of alternatively spliced forms [22,23]. One paper characterises the form that is expressed predominantly in the kidney, colon and thyroid, but lacks peptide sequence encoded by exons 3 and 7 [23]. The absence of exon 7 leads to a protein with a modified PDZ domain unable to interact with a known partner, the Mxi2 protein. Splicing-out exon 3 leads to a protein with no detectable protease activity. The authors suggest that this splice variant may have a unique role in these tissues. The genomic annotation view of the HtrA2 locus (see Fig. 1) indicates the situation may be even more complex. No less than four alternative transcripts have been deposited as mRNA entries, including additional forms that are lacking either exon 3 or exon 7. Although both the extent and functional significance of alternative splicing or polyadenylation events remains unknown, at least the availability of all potential exonic data in the genome sequence and the ability to recognise candidate polyadenylation-acceptor sites will allow these aspects of protease biology to be studied in greater depth.

7. Mutations and polymorphisms

Although mutations in individual protease loci that give rise to clinical disease are the exception rather than the rule, the distribution of these can be informative. Heamophilia B is a good example where the effects on protein domains and control regions within the factor IX serine protease locus have provided insights into the subtleties of structure, function, transcriptional regulation and phenotype [24]. No less than 689 unique molecular events have been detected in all regions of the gene except the poly (A) site. The 425 different amino acid substitutions are under-represented within the pre-peptide and activation peptide regions and over-represented in the calcium-binding EGF, and catalytic domains. The power of combining the now almost complete genome mapping data with family linkage studies is likely to reveal the molecular basis for more monogenic diseases involving proteases.

Polymorphisms can include the same kind of molecular events as mutations but at a higher population frequency (above 1%). The recent genome-wide study of 1.4 million single nucleotide polymorphisms (SNPs) shows the potential for investigating associations with complex disease traits [25]. A recent analysis of 24 kb from the ACE locus uncovered no less than 78 varying sites that could be resolved into 13 distinct haplotypes, some of which may confer an increased susceptibility to cardiovascular disease [26]. Similar levels of pop-

ulation diversity (179 variants within 66 kb) have been reported for calpain 10 [27]. Evidence suggests that certain haplotype combinations for calpain 10 may affect susceptibility to type-2 diabetes in both Mexicans and Europeans although the mechanism for such an affect is not yet clear. Those SNPs already mapped within protease loci will provide the molecular data for disease association studies [28].

8. Drug targets

The number of human proteases reported to be under investigation as drug targets more than doubled between 1998 and 2000 and now represents approximately 15% of documented proteases [29,30]. The genome data will enable a comprehensive in silico evaluation of those sequences that are already being evaluated as therapeutic targets. However, many of more recently cloned and genomically derived proteases are 'orphans' in the sense that their physiologically and/or pathologically relevant substrate(s) are unknown. Therefore, pharmaceutical or biotechnology operations should consider the industrialisation of both the expression of recombinant proteins and screening for surrogate substrates derived from peptide libraries. Such a characterisation pipeline could be prioritised by using bioinformatic filters such as signal peptides and degrees of similarity to paralogues that are already targets. Subsequent high-throughput screening for inhibitors without pre-existing target indications would be a speculative commitment of resources. However, the production of tool inhibitors and the consequent opportunity for accelerated target validation is a logical way to exploit the genome data for proteases.

9. Conclusions

By accepting a short time lag, the protease aficionado can now be assured that any new sections of transcript or genomic sequence with significant sequence similarity to a known protease will be incorporated into the secondary databases. They then have access to an extensive range of annotated features provided by genome viewers and protein domain databases. It is certain that experimental work and homology searches of genome data will continue to discover new proteases, possibly with novel mechanisms, structures or domain combinations. Numerically, however, the data suggest that the majority of human proteases are already represented in transcript data. The recent expansion of the protease collection in silico has exacerbated the bottleneck for experimental characterisation. It will be of interest to see which of the emerging technologies for high-throughput biology will provide the necessary data and expedite the conversion of orphan proteases into new drug targets.

References

[1] Clements, J., Hooper, J., Dong, Y. and Harvey, T. (2001) J. Biol. Chem. 382, 5–14.
[2] Hussain, I., Powell, D., Howlett, D.R., Tew, D.G., Meek, T.D., Chapman, C., Gloger, I.S., Murphy, K.E., Southan, C.D., Ryan, D.M., Smith, T.S., Simmons, D.L., Walsh, F.S., Dingwall, C. and Christie, G. (1999) Mol. Cell. Neurosci. 6, 419–427.
[3] Hussain, I., Powell, D.J., Howlett, D.R., Chapman, G.A., Gilmour, L., Murdock, P.R., Tew, D.G., Meek, T.D., Chapman, C., Schneider, K., Ratcliffe, S.J., Tattersall, D., Testa, T.T., Southan, C., Ryan, D.M., Simmons, D.L., Walsh, F.S., Dingwall, C. and Christie, G. (2000) Mol. Cell. Neurosci. 16, 609–619.
[4] Tipnis, S.R., Hooper, N.M., Hyde, R., Karran, E., Christie, G. and Turner, A.J. (2000) J. Biol. Chem. 275, 33238–33243.
[5] Sauder, J.M. and Arthur, J.W. (2000) J. Mol. Biol. 300, 241–248.
[6] Coates, D., Isaac, R.E., Cotton, J., Siviter, R., Williams, T.A., Shirras, A., Corvol, P. and Dive, V. (2000) Biochemistry 39, 8963–8969.
[7] Rawlings, N.D. and Barrett, A.J. (2000) Nucleic Acids Res. 28, 323–325.
[8] Pruitt, K.D. and Maglott, D.R. (2001) Nucleic Acids Res. 29, 137–140.
[9] Bairoch, A. and Apweiler, R. (2000) Nucleic Acids Res. 28, 45–48.
[10] Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M. and Servant, F. (2001) Nucleic Acids Res. 29, 37–40.
[11] International Human Genome Sequencing Consortium, (2001) Nature 409, 860–921.
[12] Southan, C. (2000) Yeast 17, 327–334.
[13] Makarova, K.S., Aravind, L. and Koonin, E.V. (2000) Trends Biochem. 25, 50–52.
[14] Southan, C. (2000) J. Pept. Sci. 6, 453–458.
[15] Mahon, P. and Bateman, A. (2000) Protein Sci. 10, 1930–1934.
[16] Caughey, G.H., Raymond, W.W., Blount, J.L., Hau, L.W., Pallaoro, M., Wolters, P.J. and Verghese, G.M. (2000) J. Immunol. 164, 6566–6575.
[17] Tatnell, P.J., Cook, M., Peters, C. and Kay, J. (2000) Eur. J. Biochem. 267, 6921–6930.
[18] Gravely, B.R. (2001) Trends Genet. 17, 100–107.
[19] Kessler, S.P., Rowe, T.M., Gomos, J.B., Kessler, P.M. and Sen, G.C. (2000) J. Biol. Chem. 275, 2659–2664.
[20] Bodendorf, U., Fischer, F., Bodian, D., Multhaup, G. and Paganetti, P. (2000) J. Biol. Chem. 276 (15), 12019–12023.
[21] Southan, C. (2000) Biochem. Soc. Trans. (abstr. no. 63) 28, 84.
[22] Gray, C.W., Ward, R.V., Karran, E., Turconi, S., Rowles, A., Viglienghi, D., Southan, C., Barton, A., Fantom, K.G., West, A., Savopoulos, J., Hassan, N.J., Clinkenbeard, H., Hanning, C., Amegadzie, B., Davis, J.B., Dingwall, C., Livi, G.P. and Creasy, C.L. (2000) Eur. J. Biochem. 267, 5699–5710.
[23] Faccio, L., Fusco, C., Viel, A. and Zervos, A.S. (2000) Genomics 68, 343–347.
[24] Lillicrap, D. (1998) Haemophilia 4 (4), 350–357.
[25] Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. and Attshuler, D. (2001) Nature 409, 928–933.
[26] Rieder, M.J. (1999) Nat. Genet. 22, 59–62.
[27] Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E., del Bosque-Plata, L., Horikawa, Y., Oda, Y., Yoshiuchi, I., Colilla, S., Polonsky, K.S., Wei, S., Concannon, P., Iwasaki, N., Schulze, J., Baier, L.J., Bogardus, C., Groop, L., Boerwinkle, E., Hanis, C.L. and Bell, G.I. (2000) Nat. Genet. 26, 163–175.
[28] Grey, I.C. (2000) Hum. Mol. Genet. 9, 2403–2408.
[29] Beeley, L.J. and Duckworth, D.M. (2000) Prog. Med. Chem. 37, 1–43.
[30] Southan, C. (2001) Drug Discov. Today, in press.