

Research article

Open Access

Identification and correction of abnormal, incomplete and mispredicted proteins in public databases

Alinda Nagy, Hédi Hegyi, Krisztina Farkas, Hedvig Tordai, Evelin Kozma, László Bányai and László Patthy*

Address: Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1113 Budapest, Hungary

Email: Alinda Nagy - nagy@enzim.hu; Hédi Hegyi - hegyi@enzim.hu; Krisztina Farkas - farkask@enzim.hu;

Hedvig Tordai - tordaih@enzim.hu; Evelin Kozma - kozma@enzim.hu; László Bányai - banyai@enzim.hu; László Patthy* - patthy@enzim.hu

* Corresponding author

Published: 27 August 2008

Received: 23 April 2008

BMC Bioinformatics 2008, 9:353 doi:10.1186/1471-2105-9-353

Accepted: 27 August 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/353>

© 2008 Nagy et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Despite significant improvements in computational annotation of genomes, sequences of abnormal, incomplete or incorrectly predicted genes and proteins remain abundant in public databases. Since the majority of incomplete, abnormal or mispredicted entries are not annotated as such, these errors seriously affect the reliability of these databases. Here we describe the MisPred approach that may provide an efficient means for the quality control of databases. The current version of the MisPred approach uses five distinct routines for identifying abnormal, incomplete or mispredicted entries based on the principle that a sequence is likely to be incorrect if some of its features conflict with our current knowledge about protein-coding genes and proteins: (i) conflict between the predicted subcellular localization of proteins and the absence of the corresponding sequence signals; (ii) presence of extracellular and cytoplasmic domains and the absence of transmembrane segments; (iii) co-occurrence of extracellular and nuclear domains; (iv) violation of domain integrity; (v) chimeras encoded by two or more genes located on different chromosomes.

Results: Analyses of predicted EnSEMBL protein sequences of nine deuterostome (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Monodelphis domestica*, *Gallus gallus*, *Xenopus tropicalis*, *Fugu rubripes*, *Danio rerio* and *Ciona intestinalis*) and two protostome species (*Caenorhabditis elegans* and *Drosophila melanogaster*) have revealed that the absence of expected signal peptides and violation of domain integrity account for the majority of mispredictions. Analyses of sequences predicted by NCBI's GNOMON annotation pipeline show that the rates of mispredictions are comparable to those of EnSEMBL. Interestingly, even the manually curated UniProtKB/Swiss-Prot dataset is contaminated with mispredicted or abnormal proteins, although to a much lesser extent than UniProtKB/TrEMBL or the EnSEMBL or GNOMON-predicted entries.

Conclusion: MisPred works efficiently in identifying errors in predictions generated by the most reliable gene prediction tools such as the EnSEMBL and NCBI's GNOMON pipelines and also guides the correction of errors. We suggest that application of the MisPred approach will significantly improve the quality of gene predictions and the associated databases.

Background

Seven years after the first drafts of the human genome were published [1,2] and three years after the completion of sequencing the human genome, the exact number of the protein-coding genes encoded in this genome is still unknown: the most likely estimates range between 20,000–25,000 genes [3,4]. More significantly, recent analyses have shown that the exact genomic structure of human protein-coding genes is correctly predicted for only about 50–60 % of the genes [5,6]. In other words, despite significant advances in computational gene identification, correct prediction of the genomic structure of the protein-coding genes of higher eukaryotes is still a very difficult task.

The main objective of our MisPred project is to develop tools that can be used to identify mispredicted genes/proteins, primarily from Metazoan genomes, in order to inform scientists of the reliability of predictions and to improve the quality of predictions. The key question is: are there signs that may indicate that the predicted structure of a protein-coding gene might be erroneous? The MisPred approach is based on the principle that a protein-coding gene is likely to be mispredicted if some of its features (or features of the protein it encodes) conflict with our current knowledge about protein-coding genes and proteins.

As a proof of principle, in the present work we describe five approaches – based on five dogmas – to identify suspicious proteins that are likely to be abnormal or mispredicted. Accordingly, the current version of MisPred contains five routines, each focusing on a special type of conflict with one of the dogmas.

(1) Conflict with the dogma that the subcellular localization of extracellular and transmembrane proteins is defined by the presence of appropriate sequence signals. The validity of this dogma is supported by studies on various disease-causing mutations, indicating that the absence of functional signal peptides prevents the translocation of proteins across the endoplasmic reticulum membrane and the mislocalized protein is rapidly degraded by the proteasome [7-13]. Similarly, the loss of functional transmembrane helices is known to lead to the mislocalization of membrane proteins [14,15]. A major reason for the rapid degradation of mislocalized extracellular proteins is that their extracellular (usually disulphide-bonded) domains are misfolded in the reductive milieu of the cytoplasm and are recognized and degraded by the protein quality control system of the cell [16].

In the current version of MisPred we used domain families the members of which occur only in the extracellular space (e.g. in secreted extracellular proteins and in the

extracellular parts of type I, type II, type III single pass or multispinning transmembrane proteins) to identify proteins that are completely or partially extracellular. The justification for this approach is that certain (usually disulphide-rich) domain families are known to have adapted (and are restricted) to the extracellular space: they occur exclusively in extracellular proteins or extracytoplasmic parts of transmembrane proteins [17,18]. Following identification of extracellular or transmembrane proteins containing such extracellular 'marker' domains, we examined whether they have the sequence signals (secretory signal peptides, signal anchors and/or transmembrane helices) that could target these domains to the extracellular space.

According to this dogma, proteins that contain obligatory extracellular domains but lack appropriate sequence signals (signal peptide, signal anchor and transmembrane segments) are considered suspicious (abnormal and non-viable) since their obligatory extracellular domains are not delivered to the extracellular space where they are stable and properly folded.

(2) Conflict with the dogma that transmembrane proteins containing both cytoplasmic and extracellular parts have at least one transmembrane segment that passes through the cell membrane. In the current version of MisPred, we used protein domain families the members of which occur exclusively in the extracellular space and exclusively in the cytoplasmic space to identify transmembrane proteins and we asked whether these proteins possess regions that pass through the cell membrane. According to this dogma, proteins that contain both obligatory extracellular and obligatory cytoplasmic domains but lack transmembrane segment(s) separating them are considered erroneous.

(3) Conflict with the dogma that obligatory extracellular and obligatory nuclear domains do not co-occur in a single, multidomain protein [17,18]. According to this dogma, proteins that contain both obligatory extracellular and obligatory nuclear domains are considered abnormal and nonviable since they cannot be delivered to a cellular compartment where both types of domains would be correctly folded and fully functional.

(4) Conflict with the rule that the protein fold is highly conserved in a domain family, therefore the number of amino acid residues in closely related members of a globular domain family usually fall into a relatively narrow range [19,20]. This phenomenon reflects the fact that the highly cooperative, rapid folding of protein domains is the result of natural selection [18,21], therefore insertion/deletion of larger segments into/from protein domains may yield macromolecules that are unable

Table 1: MisPred analysis of Swiss-Prot entries

UniProtKB/Swiss-Prot											
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	15638	1431	9.2%	15	1.05%	10	0.70%	5	0.35%	4	1
Mus musculus	13186	1198	9.1%	12	1.00%	7	0.58%	5	0.42%	2	3
Rattus norvegicus	6043	599	9.9%	18	3.01%	2	0.33%	16	2.67%	14	2
Gallus gallus	1635	194	11.9%	22	11.34%	3	1.55%	19	9.79%	12	7
Danio rerio	1290	64	5.0%	4	6.25%	3	4.69%	1	1.56%	1	0
Caenorhabditis elegans	2999	119	4.0%	9	7.56%	1	0.84%	8	6.72%	0	8
Drosophila melanogaster	2463	147	6.0%	5	3.40%	3	2.04%	2	1.36%	1	1
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	15638	43	0.3%	8	18.6%	8	18.6%	0	0.0%	0	0
Mus musculus	13186	42	0.3%	6	14.3%	6	14.3%	0	0.0%	0	0
Rattus norvegicus	6043	19	0.3%	2	10.5%	2	10.5%	0	0.0%	0	0
Gallus gallus	1635	10	0.6%	1	10.0%	1	10.0%	0	0.0%	0	0
Danio rerio	2999	2	0.1%	0	0.0%	0	0.0%	0	0.0%	0	0
Caenorhabditis elegans	1290	5	0.4%	1	20.0%	1	20.0%	0	0.0%	0	0
Drosophila melanogaster	2463	8	0.3%	1	12.5%	1	12.5%	0	0.0%	0	0
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	15638			0	0.0%	0	0.0%	0	0.0%	0	0
Mus musculus	13186			0	0.0%	0	0.0%	0	0.0%	0	0
Rattus norvegicus	6043			0	0.0%	0	0.0%	0	0.0%	0	0
Gallus gallus	1635			0	0.0%	0	0.0%	0	0.0%	0	0
Danio rerio	2999			0	0.0%	0	0.0%	0	0.0%	0	0
Caenorhabditis elegans	1290			0	0.0%	0	0.0%	0	0.0%	0	0
Drosophila melanogaster	2463			0	0.0%	0	0.0%	0	0.0%	0	0
Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	15638	6973	44.6%	10	0.14%	6	0.09%	4	0.06%	3	1
Mus musculus	13186	5808	44.0%	3	0.05%	2	0.03%	1	0.02%	1	0
Rattus norvegicus	6043	2756	45.6%	14	0.51%	0	0.00%	14	0.51%	13	1
Gallus gallus	1635	755	46.2%	8	1.06%	0	0.00%	8	1.06%	8	0
Danio rerio	1290	355	27.5%	1	0.28%	0	0.00%	1	0.28%	1	0
Caenorhabditis elegans	2999	1215	40.5%	2	0.16%	0	0.00%	2	0.16%	0	2
Drosophila melanogaster	2463	1203	48.8%	0	0.00%	0	0.00%	0	0.00%	0	0

Table 1: MisPred analysis of Swiss-Prot entries (Continued)

Conflict 5	Number of proteins	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	15638	5	0.03%	3	0.02%	2	0.01%	0	2
Mus musculus	13186	0	0.00%	0	0.00%	0	0.00%	0	0
Rattus norvegicus	6043	5	0.08%	3	0.05%	2	0.03%	0	2
Gallus gallus	1635	0	0.00%	0	0.00%	0	0.00%	0	0
Danio rerio	1290	18	1.40%	18	1.40%	0	0.00%	0	0
Caenorhabditis elegans	2999	0	0.00%	0	0.00%	0	0.00%	0	0
Drosophila melanogaster	2463	0	0.00%	0	0.00%	0	0.00%	0	0

*Values for suspicious, false positive and true positive sequences are expressed as percentage of the proteins relevant for the given conflict.

to rapidly adopt a correctly folded, viable and stable three-dimensional structure. Accordingly, proteins containing domains that consist of a significantly larger or smaller number of residues than closely related members of the same family may be suspected to be abnormal and nonviable.

(5) Conflict with the dogma that a protein is encoded by exons located on a single chromosome. According to this dogma, chimeric proteins whose parts are encoded by two or more different genes located on distinct chromosomes are considered abnormal.

In the present work we describe the results of MisPred analyses of various public databases and discuss the values and limitations of the MisPred approach.

Results and discussion

Validation of the MisPred approach on the Swiss-Prot section of the UniProtKB

The Swiss-Prot section of UniProtKB is the gold standard of protein databases therefore we have used Swiss-Prot as the benchmark with which to validate the concepts behind the MisPred approach. In view of the high quality of this manually curated database our original expectation was that very few, if any, of the Swiss-Prot entries are truly erroneous therefore it would provide a useful dataset with which to test the specificity of the different MisPred routines.

MisPred analyses of human, mouse, rat, chick, zebrafish, worm and fly Swiss-Prot entries have indeed identified very few Swiss-Prot entries as truly erroneous (see Table 1). The details of the analyses of the Swiss-Prot entries are described in Additional file 1 [see Additional file 1] and the list of the erroneous entries is deposited in Additional file 2 [see Additional file 2]. The majority of these errors could be corrected by targeted search of genomic and EST databases; the protocol used for the correction of errors will be described in another publication (manuscript in preparation).

The majority of truly erroneous sequences were returned for Conflicts 1 and 4, however, these accounted for only 0.03–1.16% and 0.008–0.49% of the sequences of the different species, respectively.

There were three major types of true positives among the Swiss-Prot entries identified by Conflict 1:

1) Fragments of full-length proteins that are not known to be fragments and/or are not annotated as such in the database. For example, LPLC4_HUMAN [Swiss-Prot:P59827] proved to be a fragment and its missing signal peptide could be predicted with the help of the sequence of a full-length mouse ortholog (Figure 1). Similarly, the sequence of C209C_MOUSE [Swiss-Prot:Q91ZW9] (lacking a transmembrane helix) could be corrected by targeted search of mouse genomic and EST sequences (Figure 2).

2) Mispredicted proteins. The hypothetical worm protein YL15_CAEL [Swiss-Prot:Q11101] is an example for this type of error. The protein arose through *in silico* fusion of a gene related to the homeobox protein HM07_CAEL [Swiss-Prot:P20270] and a gene related to the Kunitz_BPTI containing protein CBG14258, Q619J1_CAEBR [TrEMBL:Q619J1] (Figure 3).

3) Proteins translated from aberrant transcripts that do not encode viable proteins. For example NOE2_MOUSE [Swiss-Prot:Q8BM13] lacks a signal peptide, whereas the rat ortholog [RefSeq:NP_001015017] and a different isoform of this mouse protein [GenBank:EDL25126] do possess a signal sequence.

The MisPred routine used for the detection of Conflict 1 is characterized by a very low number of false positives. There are three main sources of false positives:

1) Some proteins are identified as suspicious due to the limitations of the bioinformatic tools incorporated in this MisPred routine (e.g. failure to detect some signal peptides and transmembrane helices).

	1	50
lplc4_human_corrected	<u>MWMAWCVAALS</u> <u>VVAVCG</u> TSHETNTVLRVTKDVLNSAISGMLQQSDALHSA	
lplc4_human	~~~~~MLQQSDALHSA	
CAM20161	<u>MWTAWCVAALS</u> <u>VAAVCG</u> IRQDTTTLRVTKDVLGNAISGTIQKSDAFRSA	
	51	100
lplc4_human_corrected	LREVPLGVGDIPYNDHFVRGPPPVYTNKGKLDGIYQYGHIEETNDNTAQLG	
lplc4_human	LREVPLGVGDIPYNDHFVRGPPPVYTNKGKLDGIYQYGHIEETNDNTAQLG	
CAM20161	LREVPVGVGGVPYNDHFVREPPPKYTNGRQLGGNYKYGHIAKANDNRAQLG	
	101	150
lplc4_human_corrected	GKYRYGEILESEGSIRDLRNSGYRSAENAYGGHRGLGRYRAA...PVGR	
lplc4_human	GKYRYGEILESEGSIRDLRNSGYRSAENAYGGHRGLGRYRAA...PVGR	
CAM20161	GKYRYGEILDSDGSLRDLRHEDYRPPDSAY...HRGSGRYRSAADSSSVGR	
	151	200
lplc4_human_corrected	LHRRELQPGEIPPGVATGAVGPGGLLTGGMLAADGILAGQGGLLGGGGL	
lplc4_human	LHRRELQPGEIPPGVATGAVGPGGLLTGGMLAADGILAGQGGLLGGGGL	
CAM20161	LYRRELRPGEIPAGVATGALGPGGLLTGGMLANEGILAGQGGLLGGGGL	
	201	250
lplc4_human_corrected	LGDGGLLGGGGVLGVLGEGGILSTVQGITGLRIVELTLPRVSVRLLPVG	
lplc4_human	LGDGGLLGGGGVLGVLGEGGILSTVQGITGLRIVELTLPRVSVRLLPVG	
CAM20161	LGDGGLLGGGGVLGVLGEGGILSTVQGITGLRIVELTLPRVSVRLLPVG	

Figure 1

Error detected by MisPred routine for Conflict 1: the case of the Swiss-Prot entry LPLC4_HUMAN. The protein contains extracellular domains LBP_BPI_CETP and LBP_BPI_CETP_C but was found to lack both a signal peptide and trans-membrane helices. The human sequence was corrected (LPLC4_HUMAN_corrected) by targeted search of the human genome with its mouse ortholog, CAM20161 [EMBL:CAM20161] that has a signal peptide. The alignment shows the N-terminal parts of LPLC4_HUMAN, CAM20161 and LPLC4_HUMAN_corrected. The predicted signal peptides of CAM20161 and LPLC4_HUMAN_corrected are in yellow and underlined.

2) Exceptions to the dogma on which this MisPred routine is based, i.e. some secreted proteins truly lack secretory signal peptides since they are subject to leaderless protein secretion [22], such as the secreted proteins GAPR1_HUMAN [Swiss-Prot:Q9H4G4] and TINAG_HUMAN [Swiss-Prot:Q9UIW2].

3) Exceptions to the rule that all members of an extracellular domain family are restricted to the extracellular space.

MisPred routine based on Conflict 4 also identified a number of truly erroneous entries. A major source for this type of error is that the Swiss-Prot entry corresponds to an incomplete protein (with a truncated domain). For example, the sequence of EPHA5_RAT [Swiss-Prot: P54757] contains only a fragment of a SAM_1 domain since the protein sequence is truncated at the C-terminal end (Fig-

ure 4). The error in EPHA5_RAT could be corrected by targeted search of the rat genome using the sequences of the full-length orthologs (see Figure 4).

The routine based on Conflict 5 identified no erroneous mouse, chicken, worm or fruitfly Swiss-Prot proteins and the proportion of trans-chromosomal chimeras is very low in the case of human (0.01%) and rat (0.03%) sequences. On the other hand, the frequency of suspected chimeras is relatively high in the case of zebrafish (1.4%). As discussed in Additional file 1 [see Additional file 1], the most likely explanation for this observation is that they are false positives: the chromosomal assignment of contigs encoding different parts of these zebrafish genes may not be correct.

MisPred analyses have revealed that none of the Swiss-Prot entries violate the rules underlying Conflicts 2 and 3.

	1	50
c209c_mouse	~~~~~	~~~~~
c209c_mouse_corrected	MNDSAEGRVQQLGSLDEEHLIPSGTRYSFKGFRFQSPYVLKKTAGYLCHG	
c209a_mouse	MSDSKEMGKRQLRPLDEELLTSSHTRHSIKGFGFQTNSGFSSFTGCLVHS	
c209d_mouse	MSDSMESKTQQVVIPEDEECLMSGTRYSDISSRLQTKFGIKSLAEYTKQS	
	51	100
c209c_mouse	~~~~~MRMHTRLQFLKRVSNAVYSHGQEQAKKEKVYKEMTQLK	
c209c_mouse_corrected	LGP <u>FVLQFLC~~LT~~STILLAILV</u> KVSNVAYSHGQEQAKKEKVYKEMTQLK	
c209a_mouse	QVP <u>LALQVLFLAVCSVLLVVILV</u> KVYKIPSS..QEENNQMNVYQELTQLK	
c209d_mouse	RNP <u>LVLQLLSFLFLAGLLLIILILV</u> SKVPSSEVQN.....KIYQELMQLK	
	101	150
c209c_mouse	SQI.NRLCRPCPWDWTVFQGNCYFFSKFQQNWNSVNACRKLDAQLVVIK	
c209c_mouse_corrected	SQI.NRLCRPCPWDWTVFQGNCYFFSKFQQNWNSVNACRKLDAQLVVIK	
c209a_mouse	AGV.DRLCRSCPWDWTHFQGSYFFSVAQKSWNDSATACHNVGAQLVVIK	
c209d_mouse	AEVHDGLCQPCARDWTFNGSCYFFSKSQRNWHNSTTACQELGAQLVIE	
	151	200
c209c_mouse	SDDEQSFLQQTSKEKGYAWMGLSDLKHEGRWHWVDGSHLLFSFMKYWNKG	
c209c_mouse_corrected	SDDEQSFLQQTSKEKGYAWMGLSDLKHEGRWHWVDGSHLLFSFMKYWNKG	
c209a_mouse	SDEEQNFLQQTSKKRGYTWMGLIDMSKESTWYVVDGSPLTLSFMKYWSKG	
c209d_mouse	TDEEQTFLLQQTSKARGPTWMGLSDMHNEATWHWVDGSPLSPSFTRYWNRG	
	201	242
c209c_mouse	EPNNEWEEDCAEFRGDGWNDA ~~PCT~~ IKKYWICKK.SAMSCTEK	
c209c_mouse_corrected	EPNNEWEEDCAEFRGDGWNDA ~~PCT~~ IKKYWICKK.SAMSCTEK	
c209a_mouse	EPNNLGEEDCAEFRDDGWN ~~DTK~~ CTNKKFWICKK.LSTSCPSK	
c209d_mouse	EPNNVGDEDCAEFSGDGWN ~~DL~~ SCDKLLFWICKKVSTSSCTTK	

Figure 2

Error detected by MisPred routine for Conflict 1: the case of the Swiss-Prot entry C209C_MOUSE. The protein contains an extracellular C-type lectin domain but was found to lack both a signal peptide and transmembrane helices, whereas all closely related proteins (e.g. C209A_MOUSE, C209D_MOUSE [Swiss-Prot:Q91ZXI, Q91ZW8]) are type II transmembrane proteins. The sequence of this protein was corrected by targeted search of mouse genomic and EST sequences. The alignment shows the N-terminal parts of C209C_MOUSE, C209C_MOUSE_corrected, C209A_MOUSE and C209D_MOUSE. The predicted transmembrane helices of C209C_MOUSE_corrected, C209A_MOUSE and C209D_MOUSE are in red and underlined.

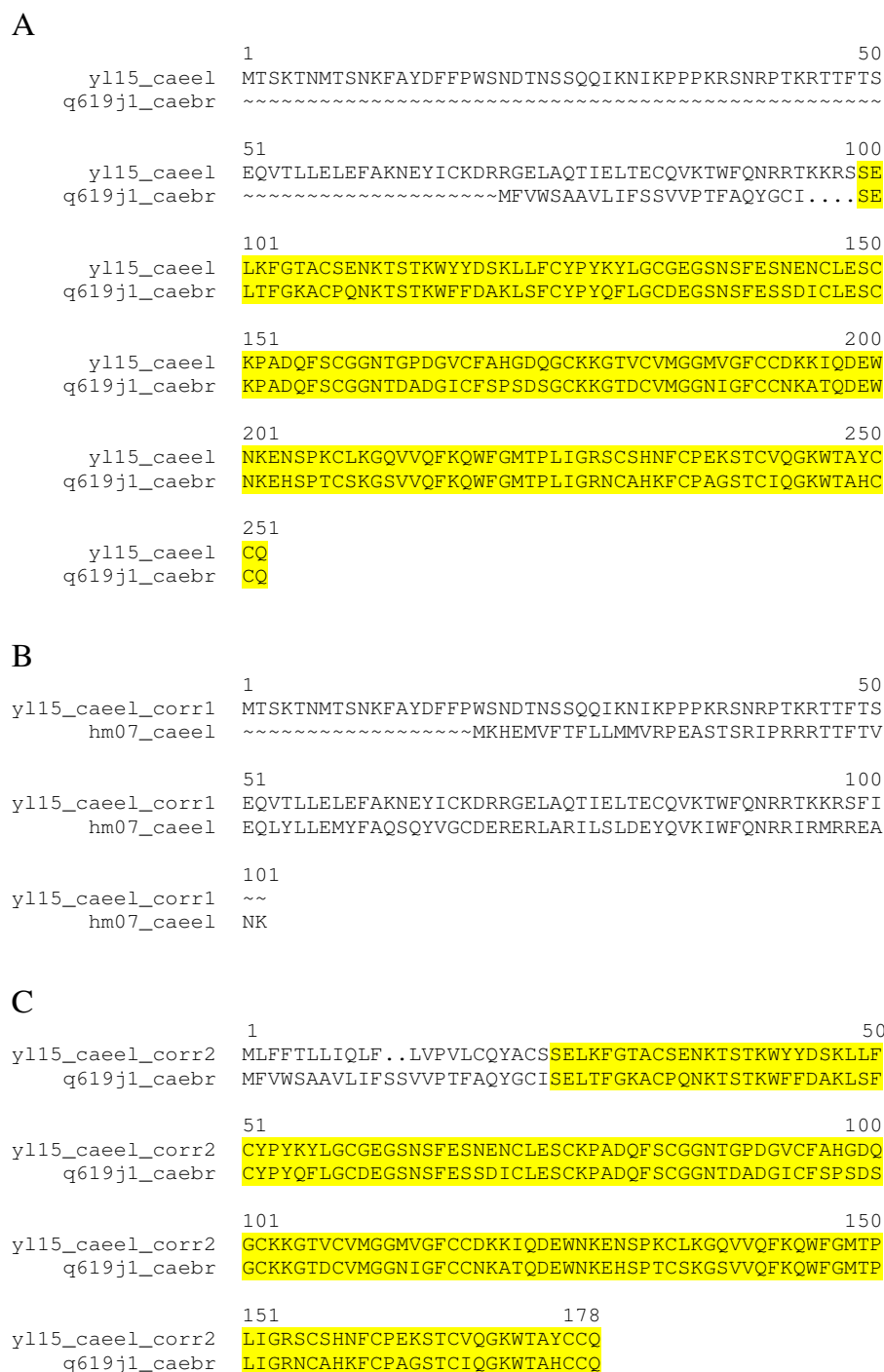
The fact that the number of Swiss-Prot entries identified by MisPred as erroneous is very low attests to both the high quality of this database and the reliability of the MisPred approach. Assuming that the vast majority of the Swiss-Prot sequences that were not detected by MisPred routines are true negatives the false positive rate of the routines was calculated to be ≤ 0.001 , i.e. their specificity is very high (≥ 0.999).

MisPred analysis of the TrEMBL section of UniProtKB

The primary motivation for MisPred analysis of UniProtKB/TrEMBL was that TrEMBL entries are used in vari-

ous types of evidence-based, extrinsic gene prediction programs and thus have a strong influence on the quality of gene predictions. The results of the analyses of human proteins are summarized in Table 2.

The data shown in Table 2 indicate that the proportion of suspicious TrEMBL entries is relatively high in the case of Conflict 1, Conflict 4 and Conflict 5. Importantly, these values are orders of magnitude higher than those for the Swiss-Prot entries, indicating that the vast majority of TrEMBL proteins identified by MisPred as suspicious are truly erroneous.

**Figure 3**

Error detected by MisPred routine for Conflict 1: the case of the Swiss-Prot entry YL15_CAEEL The hypothetical homeobox protein C02F12.5 [EnsEMBL: C02F12.5] predicted for chromosome X contains an extracellular Kunitz_BPTI domain but was found to lack both a signal peptide and transmembrane helices. This protein, that also contains a nuclear Homeobox domain, arose through *in silico* fusion of a gene related to the homeobox protein HM07_CAEEL and a gene related to the Kunitz_BPTI containing protein CBG14258, Q619J1_CAEBR. (A) Alignment of YL15_CAEEL and Q619J1_CAEBR shows close homology only in the C-terminal region, highlighted in yellow. (B) Alignment of the YL15_CAEEL_corr1 and HM07_CAEEL. (C) Alignment of YL15_CAEEL_corr2 and Q619J1_CAEBR.

	751	800
epha5_rat_corrected	PVMIVTEYMENGSLDTFLKKNDGQFTVIQLVGMLRGIAAGMKYLSDMGYV	
epha5_rat	PVMIVTEYMENGSLDTFLKKNDGQFTVIQLVGMLRGIAAGMKYLSDMGYV	
epha5_human	PVMIVTEYMENGSLDTFLKKNDGQFTVIQLVGMLRGISAGMKYLSDMGYV	
epha5_chick	PVMIVTEYMENGSLDTFLKKNDGQFTVIQLVGMLRGISAGMKYLSDMGYV	
epha5_mouse	PVMIVTEYMENGSLDTFLKKNDGQFTVIQLVGMLRGIAAGMKYLSDMGYV	
	801	850
epha5_rat_corrected	HRDLAARNILINSNLVCKVSDFGLSRVLEDDPEAAAYTTRGGKIPIRWTAP	
epha5_rat	HRDLAARNILINSNLVCKVSDFGLSRVLEDDPEAAAYTTRGGKIPIRWTAP	
epha5_human	HRDLAARNILINSNLVCKVSDFGLSRVLEDDPEAAAYTTRGGKIPIRWTAP	
epha5_chick	HRDLAARNILINSNLVCKVSDFGLSRVLEDDPEAAAYTTRGGKIPIRWTAP	
epha5_mouse	HRDLAARNILINSNLVCKVSDFGLSRVLEDDPEAAAYTTRGGKIPIRWTAP	
	851	900
epha5_rat_corrected	EAIAFRKFTSASDVWSYGIVMWEVVS YGERPYWEMTNQDVIKAVEEGYRL	
epha5_rat	EAIAFRKFTSASDVWSYGIVMWEVVS YGERPYWEMTNQDVIKAVEEGYRL	
epha5_human	EAIAFRKFTSASDVWSYGIVMWEVVS YGERPYWEMTNQDVIKAVEEGYRL	
epha5_chick	EAIAFRKFTSASDVWSYGIVMWEVMS YGERPYWEMTNQDVIKAVEEGYRL	
epha5_mouse	EAIAFRKFTSASDVWSYGIVMWEVVS YGERPYWEMTNQDVIKAVEEGYRL	
	901	950
epha5_rat_corrected	PSPMDCPAALYQLMLDCWQKDRNSRPKFDDIVNMLDKLIRNPSSLKTLVN	
epha5_rat	PSPMDCPAALYQLMLDCWQKDRNSRPKFDDIVNMLDKLIRNPSSLKTLVN	
epha5_human	PSPMDCPAALYQLMLDCWQKERNRPKFDEIVNMLDKLIRNPSSLKTLVN	
epha5_chick	PSPMDCPAALYQLMLDCWQKDRNSRPKFDEIVSMLDKLIRNPSSLKTLVN	
epha5_mouse	PSPMDCPAALYQLMLDCWQKDRNSRPKFDEIVNMLDKLIRNPSSLKTLVN	
	951	1000
epha5_rat_corrected	ASSRVSTLLAEHGSLGSGAYRSVGEWLEATKMGRYTEIFMENGYSSMDAV	
epha5_rat	ASSRVSTLLAEHGSLGSGAYRSVGEWLEATKMGRYTEIFMENGYSSMDAV	
epha5_human	ASCRVSNLLAEHSPLGSGAYRSVGEWLEAIKMGRYTEIFMENGYSSMDAV	
epha5_chick	ASSRVSNLLVEHSPVGSGAYRSVGEWLEAIKMGRYTEIFMENGYSSMDSV	
epha5_mouse	ASSRVSTLLAEHGSLGSGAYRSVGEWLEAIKMGRYTEIFMENGYSSMDAV	
	1001	1041
epha5_rat_corrected	AQVTLE <u>DLRRLGVTLVGHQ.KKIMNSLQEMKVQLVNGMVPV</u>	
epha5_rat	AQVTLE~~~~~	
epha5_human	AQVTLEDLRRLGVTLVGHQ.KKIMNSLQEMKVQLVNGMVPL	
epha5_chick	AQVTLEDLRRLGVTLVGHQ.KKIMNSLQEMKVQLVNGMVPL	
epha5_mouse	AQVTLEDLRRLGVTLVGHQKKKIMSSLQEMKVQMVNGMVPV	

Figure 4

Error detected by MisPred routine for Conflict 4: the case of the Swiss-Prot entry EPHA5_RAT. This protein contains a C-terminal truncated SAM_I domain that deviates significantly from the normal size of this domain family. It is noteworthy that orthologs from mouse, human and chicken contain an intact SAM_I domain. The sequence of this protein was corrected by targeted search of the rat genome using the sequences of the full-length orthologs. The alignment shows the C-terminal parts of EPHA5_RAT, EPHA5_RAT_corrected, EPHA5_MOUSE [Swiss-Prot:Q60629], EPHA5_HUMAN [Swiss-Prot:P54756] and EPHA5_CHICK [Swiss-Prot:P54755]. The region of the predicted SAM_I domain of EPHA5_RAT_corrected that is absent in EPHA5_RAT is underlined and highlighted in yellow.

Table 2: MisPred analysis of TrEMBL entries

UniProtKB/TrEMBL											
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	52237	6732	12.9%	3907	58.0%	ND	ND	ND	ND	ND	ND
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	52237	58	0.11%	9	15.5%	9	15.5%	0	0.0%	0	0
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	52237			0	0.0%	0	0.0%	0	0.0%	0	0
Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	52237	17073	32.7%	2531	14.8%	ND	ND	ND	ND	ND	ND
Conflict 5	Number of proteins			Identified as suspicious by MisPred	Percentage*	False positives	Percentage*	True errors	Percentage*	Annotated as fragment or chimera by UniProt	Identified as abnormal only by MisPred
Homo sapiens	52237			172	0.33%	0	0.00%	172	0.33%	85	87
Mus musculus	50304			40	0.08%	ND	ND	ND	ND	ND	ND
Rattus norvegicus	8557			5	0.06%	ND	ND	ND	ND	ND	ND
Gallus gallus	5549			6	0.11%	ND	ND	ND	ND	ND	ND
Danio rerio	19623			387	1.97%	ND	ND	ND	ND	ND	ND
Caenorhabditis elegans	30000			0	0.00%	0	0	0	0	0	0
Drosophila melanogaster	26947			49	0.18%	ND	ND	ND	ND	ND	ND

*Values for suspicious, false positive and true positive sequences are expressed as percentage of the proteins relevant for the given conflict.
 ND – not determined

The majority (58.0%) of human TrEMBL proteins containing at least one extracellular domain were found by MisPred to lack a signal peptide and/or a transmembrane segment in contrast to 1.05% in the case of human Swiss-Prot entries. Similarly, 14.8% of human TrEMBL entries containing at least one member of the Pfam-A domain families suitable for the study of domain integrity were found to contain a domain of abnormal size, while this value is only 0.14% in Swiss-Prot. The reason why a high proportion of TrEMBL proteins are identified by Conflict 1 as suspicious is that many TrEMBL entries are truncated at the N-terminal end and N-terminally truncated secreted proteins are likely to lack the signal peptides. Similarly,

the high proportion of TrEMBL entries affected by Conflict 4 reflects the severe contamination of this database with proteins predicted for incomplete cDNAs. Since cDNAs are more likely to be incomplete at their 5' end than their 3' end, the size of Pfam-A domains at the N-terminal end of proteins of the TrEMBL database was found to deviate more significantly from the average size than those of internal domains (data not shown), again indicating that a relatively large proportion of TrEMBL entries are truncated at the N-terminal end. In harmony with this explanation, 95% (Conflict 1) and 100% (Conflict 4) of the suspicious entries are also annotated as fragments in TrEMBL.

A

	951	1000
prd10_human	<u>QQQNSSVQHTYLPSAWNSFRGYSSEIQMMTLPPGQFVITDSGVATPVTTG</u>	
q9nxi4_human	<u>QQQNSSVQHTYLPSAWNSFRGYSSEIQMMTLPPGQFVITDSGVATPVTTG</u>	
	1001	1050
prd10_human	<u>QVKAVTSGHYVLSESQSELEEKQTSALSGGVQVEPPAHSDSLDPQTNSQQ</u>	
q9nxi4_human	<u>QVKAVTSGHYVLSESQSELEEKQTSALSGGVQVEPPAHSDSLDPQTNSQQ</u>	
	1051	1100
prd10_human	<u>QTTQYIIT</u> TTTNGNGSSEVHITKP~~~~~	
q9nxi4_human	<u>QTTQYIIT</u> AGSKVI.QNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTT	
	1101	1133
prd10_human	~~~~~	
q9nxi4_human	FKNIKSVTELNGDIITNTMTLGDIVFKRISKRI	

B

	701	750
fabpl_human	~~~~~	
q9nxi4_human	EPTASAPSSAQVSGQPLSPSAQQAQQLSPSHIQGSSSTQGQALQQQQQQ	
	751	800
fabpl_human	~~~~~	
q9nxi4_human	QQNSSVQHTYLPSAWNSFRGYSSEIQMMTLPPGQFVITDSGVATPVTTGQ	
	801	850
fabpl_human	~~MSFSGKYQL.QSQENFEAFMKAI...GLPEELIQKGKDIKGVSEIVQN	
q9nxi4_human	VKAVTSGHYVLSESQSELEEKQTSALSGGVQVEPPAHSDSLDPQTNSQQ.	
	851	900
fabpl_human	GKHFKFTITAGSKVIQNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTT	
q9nxi4_human	.QTTQYIITAGSKVIQNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTT	
	901	933
fabpl_human	FKNIKSVTELNGDIITNTMTLGDIVFKRISKRI	
q9nxi4_human	FKNIKSVTELNGDIITNTMTLGDIVFKRISKRI	

Figure 5

Error detected by MisPred routine for Conflict 5: the case of the protein Q9NXI4_HUMAN. The cDNA of this hypothetical protein FLJ20227, cloned from colon mucosa is derived from a chimera of two genes located on chromosome 11 and chromosome 2. The N-terminal part of the protein (underlined and highlighted in yellow) is derived from the gene encoding the PR domain zinc finger protein 10, PRD10_HUMAN (A), the C-terminal part of the protein (underlined and highlighted in blue) is derived from the gene encoding liver fatty acid-binding protein, FABPL_HUMAN (B).

Table 3: MisPred analysis of EnsEMBL entries

EnsEMBL									
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	48403	3449	7.13%	277	8.03%	ND	ND	ND	ND
Mus musculus	31302	2038	6.51%	151	7.41%	ND	ND	ND	ND
Rattus norvegicus	33745	2390	7.08%	325	13.6%	ND	ND	ND	ND
Monodelphis domestica	32690	2369	7.25%	661	27.9%	ND	ND	ND	ND
Gallus gallus	24168	1519	6.29%	413	27.19%	ND	ND	ND	ND
Xenopus tropicalis	28324	2383	8.41%	931	39.07%	ND	ND	ND	ND
Fugu rubripes	22102	1612	7.29%	627	38.9%	ND	ND	ND	ND
Danio rerio	36065	3312	9.18%	1224	36.96%	ND	ND	ND	ND
Ciona intestinalis	20000	1452	7.26%	670	46.14%	ND	ND	ND	ND
Caenorhabditis elegans	26439	918	3.47%	117	12.75%	ND	ND	ND	ND
Drosophila melanogaster	19789	1071	5.41%	120	11.2%	ND	ND	ND	ND
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	48403	101	0.21%	18	17.82%	18	17.82%	0	0.00%
Mus musculus	31302	50	0.16%	4	8.00%	4	8.00%	0	0.00%
Rattus norvegicus	33745	67	0.2%	12	17.91%	10	14.93%	2	2.99%
Monodelphis domestica	32690	101	0.31%	25	24.75%	9	8.91%	16	15.84%
Gallus gallus	24168	45	0.19%	5	11.11%	4	8.89%	1	2.22%
Xenopus tropicalis	28324	57	0.2%	11	19.3%	5	8.77%	6	10.53%
Fugu rubripes	22102	58	0.26%	19	32.76%	12	20.69%	7	12.07%
Danio rerio	36065	75	0.21%	8	10.67%	7	9.33%	1	1.33%
Ciona intestinalis	20000	29	0.15%	2	6.90%	2	6.90%	0	0.00%
Caenorhabditis elegans	26439	12	0.05%	1	8.33%	1	8.33%	0	0.00%
Drosophila melanogaster	19789	16	0.08%	1	6.25%	1	6.25%	0	0.00%
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	48403			1	0.002%	0	0.00%	1	0.002%
Mus musculus	31302			3	0.01%	0	0.00%	3	0.01%
Rattus norvegicus	33745			3	0.01%	0	0.00%	3	0.01%
Monodelphis domestica	32690			0	0.00%	0	0.00%	0	0.00%
Gallus gallus	24168			1	0.004%	0	0.00%	1	0.004%
Xenopus tropicalis	28324			0	0.00%	0	0.00%	0	0.00%
Fugu rubripes	22102			2	0.01%	0	0.00%	2	0.01%
Danio rerio	36065			0	0.00%	0	0.00%	0	0.00%

Table 3: MisPred analysis of EnsEMBL entries (Continued)

Ciona intestinalis	20000			0	0.00%	0	0.00%	0	0.00%
Caenorhabditis elegans	26439			0	0.00%	0	0.00%	0	0.00%
Drosophila melanogaster	19789			0	0.00%	0	0.00%	0	0.00%

Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	48403	16681	34.46%	850	5.1%	ND	ND	ND	ND
Mus musculus	31302	9955	31.80%	306	3.07%	ND	ND	ND	ND
Rattus norvegicus	33745	11826	35.05%	474	4.01%	ND	ND	ND	ND
Monodelphis domestica	32690	11847	36.24%	381	3.22%	ND	ND	ND	ND
Gallus gallus	24168	6261	25.91%	383	6.12%	ND	ND	ND	ND
Xenopus tropicalis	28324	6733	23.78%	318	4.72%	ND	ND	ND	ND
Fugu rubripes	22102	5464	24.72%	278	5.09%	ND	ND	ND	ND
Danio rerio	36065	9402	26.07%	591	6.29%	ND	ND	ND	ND
Ciona intestinalis	20000	2114	10.57%	147	6.95%	ND	ND	ND	ND
Caenorhabditis elegans	26439	3039	11.49%	86	2.83%	ND	ND	ND	ND
Drosophila melanogaster	19789	3341	16.88%	58	1.74%	ND	ND	ND	ND

Conflict 5	Number of proteins	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	48403	0	0.00%	0	0.00%	0	0.00%
Danio rerio	36065	9	0.02%	7	0.02%	2	0.01%

*Values for suspicious, false positive and true positive sequences are expressed as percentage of the proteins relevant for the given conflict.
 ND – not determined

Errors of TrEMBL entries are not only due to the incompleteness of cDNAs; transcripts formed through aberrant splicing and chimeric transcripts may also contribute to errors in this database. Interestingly, there are numerous human TrEMBL entries that are chimeric (0.33%), different segments of the predicted protein sequences being encoded by different genes located on different chromosomes. A large proportion (43.6%) of these chimeric entries are annotated as resulting from the fusion of genes located on different chromosomes through chromosomal translocation in a cancer cell line, 7.6% have no such annotation although the corresponding cDNAs were cloned from cancer tissues. It should be pointed out, however, that there are many chimeric proteins in UniProtKB derived from cDNAs that were cloned from apparently normal tissues (36.6%), suggesting that chimera formation is more general than previously thought. For example, the cDNA of the hypothetical protein FLJ20227 [TrEMBL:Q9NXI4], cloned from colon mucosa, is a chi-

mera of two genes located on chromosome 11 and chromosome 2. The N-terminal part of the protein is derived from the gene encoding the PR domain zinc finger protein 10 (PRD10_HUMAN) [Swiss-Prot:Q9NQV6], the C-terminal part is derived from the gene encoding the liver form of Fatty acid-binding protein (FABPL_HUMAN) [Swiss-Prot:P07148] (see Figure 5). Another factor that increased the number of human chimeric proteins in TrEMBL is that the biotechnology industry has contributed numerous synthetic (chimeric) human constructs to the TrEMBL database (5.8%).

It is also noteworthy that the rate of chimeras is much higher in the case of zebrafish sequences (1.97%) than in the case of other vertebrates. As discussed in Additional file 1 [see Additional file 1], the most likely explanation for this observation is that the chromosomal assignment of contigs encoding these zebrafish genes may not be correct: contigs carrying different fragments of a zebrafish

Table 4: MisPred analysis of NCBI's GNOMON-predicted proteins

NCBI/GNOMON									
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	10125	287	2.83%	93	32.4%	ND	ND	ND	ND
Monodelphis domestica	20110	1293	6.43%	253	19.57%	ND	ND	ND	ND
Gallus gallus	14816	909	6.14%	246	27.06%	ND	ND	ND	ND
Danio rerio	25356	2108	8.31%	562	26.66%	ND	ND	ND	ND
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	10125	4	0.04%	0	0%	0	0.00%	0	0.00%
Monodelphis domestica	20110	32	0.16%	6	18.75%	3	9.38%	3	9.38%
Gallus gallus	14816	22	0.15%	5	22.73%	3	13.64%	2	9.09%
Danio rerio	25356	31	0.12%	11	35.48%	5	16.13%	6	19.35%
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	10125			0	0.00%	0	0.00%	0	0.00%
Monodelphis domestica	20110			2	0.01%	0	0.00%	2	0.01%
Gallus gallus	14816			2	0.01%	1	0.01%	1	0.01%
Danio rerio	25356			7	0.03%	3	0.01%	4	0.02%
Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	10125	1632	16.12%	255	15.63%	ND	ND	ND	ND
Monodelphis domestica	20110	6224	30.95%	111	1.78%	ND	ND	ND	ND
Gallus gallus	14816	3564	24.06%	370	10.38%	ND	ND	ND	ND
Danio rerio	25356	4387	17.31%	385	8.78%	ND	ND	ND	ND
Conflict 5	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	10125			1	0.01%	0	0.00%	1	0.01%
Danio rerio	25356			25	0.10%	24	0.09%	1	0.004%

*Values for suspicious, false positive and true positive sequences are expressed as percentage of the proteins relevant for the given conflict.
 ND – not determined

gene may be incorrectly assigned to different chromosomes.

MisPred routines for Conflict 2 and Conflict 3 identified no TrEMBL entries as erroneous. This is primarily due to the fact that sequences predicted *in silico* (that could miss internal transmembrane segments separating extracellular and cytoplasmic domains or fuse tandem genes) are absent from this section of UniProtKB.

MisPred analysis of the EnsEMBL database and the GNOMON-predicted proteins of the NCBI database

Table 3 and Table 4 summarize the results of the analysis of the EnsEMBL database and the GNOMON-predicted proteins of the NCBI database for the various species and the five different MisPred routines (for details [see Additional file 1]).

As illustrated in Tables 3 and 4, a relatively high proportion of EnsEMBL and GNOMON-predicted entries are detected by the routine for Conflict 1 as suspicious (ranging from 8 to 46% for EnsEMBL and 20 to 32% for NCBI entries containing extracellular domains). The most likely explanation for this is that in a large proportion of secreted vertebrate proteins the signal peptide is encoded by an exon separated by a long intron from the downstream exon [23]. Low sequence conservation of signal peptides and low transcript- and EST-coverage of 5' parts of protein-coding genes explain why finding short exons encoding just the signal peptides has a rather low rate of success in the case of vertebrate genomes. This problem is less serious in the case of intron-poor genomes, such as those of *Caenorhabditis elegans* or *Drosophila melanogaster* since their signal peptides are less likely to be encoded by short, distinct exons [24]; this is reflected in a lower proportion of suspicious proteins in the case of EnsEMBL entries of these species. Furthermore, worm and fruitfly were the first metazoan organisms whose genome sequences were determined [25,26] and genome annotation efforts of nearly a decade have significantly improved the quality of gene predictions.

Very few erroneous EnsEMBL and GNOMON-predicted sequences were detected by the routine for Conflict 2. The explanation for the low rate of this type of error is that – unlike signal peptide segments – transmembrane helices are usually encoded by longer exons that also encode other conserved parts of transmembrane proteins, facilitating their detection. Another factor that facilitates detection of exons of transmembrane regions is that they are more likely to be located in the middle or 3' parts of genes whose transcript- and EST-coverage is relatively high. The true positives for Conflict 2 were found to be of two major types:

(1) The predicted protein lacks transmembrane helices since the corresponding region of the gene is mispredicted. For example, ENSXETP00000040601 [EnsEMBL: ENSXETP00000040601], which corresponds to the frog ortholog of Ephrin receptor A7, lacks a typical transmembrane helix between its extracellular and cytoplasmic domains; the missing transmembrane sequence could be corrected using frog EST sequences (Figure 6). Detailed analysis of this group of true positives (vertebrate EPH receptor tyrosine kinases, Tie-2 receptor tyrosine kinases, skeletal muscle receptor tyrosine kinases, receptor-type tyrosine-protein phosphatases, Notch proteins, etc.) revealed that the regions containing their transmembrane helices are encoded by relatively short exons distinct from those encoding conserved extracellular and cytoplasmic domains [27], making it difficult to find these exons.

(2) The gene was mispredicted by *in silico* fusion of distinct, tandem genes encoding extracellular and cytoplasmic proteins. Several examples of this type of error were found among *Fugu rubripes* proteins, but not in the case of other organisms, including zebrafish. A possible explanation for this observation is that the intergenic distance is significantly shorter in the compact genome of pufferfish than in the case of other vertebrate genomes [28], increasing the chance of *in silico* fusion of tandem genes.

MisPred routine for Conflict 3 detected very few errors in predicted proteins. Analyses of these sequences have revealed that they arose as a result of *in silico* fusion of two or more distinct, tandem genes encoding extracellular and nuclear proteins. Interestingly, proteins containing extracellular Pentaxin and nuclear Chromo domains were found among human, mouse, rat and chicken EnsEMBL proteins. There are several interpretations for their occurrence in different warm-blooded animals. One possible explanation is that since the constituent genes are closely linked in all these species, gene-prediction erroneously fused these otherwise independent genes. An alternative explanation is that these genes truly give rise to novel transcripts and proteins in which nuclear Chromo domains are fused to extracellular Pentaxin domains. In other words, nuclear Chromo domains can co-occur with the extracellular Pentaxin domains, either because the Chromo domain is not an obligatory nuclear domain or the Pentaxin domain is not an obligatory extracellular domain. It is noteworthy in this respect that Chen and Bixby [29,30] have cloned three mouse variants of neuronal pentraxin with Chromo domain (Q6TLW1_MOUSE, Q6TLW0_MOUSE, Q6TKP2_MOUSE) [Swiss-Prot:Q6TLW1, Q6TLW0, Q6TKP2] but none of them have signal peptides suggesting that, unlike the major products of the neuronal pentraxin genes, they are not secreted. It is still possible, however, that the fusion proteins encoded by chimeric transcripts are abnormal in the sense that they

ensxetp00000040601_corrected	401	450
np_990414	SIVDLVAHANYTFEVEAVNGVSDLSRSQRLFAAVSVTTGQAAPSQVSGVM	
ensxetp00000040601	TVMDLLAHANYTFEVEAVNGVSDLSRSQRLFAAVSITTGQAAPSQVSGVM	
	SIVDLVAHANYTFEVEAVNGVSDLSRSQRLFAAVSVTTGQAAPSQVSGVM	
ensxetp00000040601_corrected	451	500
np_990414	KERVLQRAVDLSWQEPEHPNGVITEYEIKYYEKDQRERTYSTLTKTKSTSV	
ensxetp00000040601	KERVLQRSVELSWQEPEHPNGVITEYEIKYYEKDQRERTYSTVTKTKSTSA	
	KERVLQRAVDLSWQEPEHPNGVITEYEIKYYEKDQRERTYSTLTKTKSTSV	
ensxetp00000040601_corrected	501	550
np_990414	SINNLRPGTAYIFQIRAFATAAGYGMYSRPLDVSTLEEATATAVSTEQNPV	
ensxetp00000040601	SINNLRPGTVYVFQIRAFATAAGYGNYSRPLDVATLEEATATAVSSEQNPV	
	SINNLRPGTAYIFQIRAFATAAGYGMYSRPLDVSTLEEATVYYIFACSYCI	
ensxetp00000040601_corrected	551	600
np_990414	<u>III</u> AVVAVAGT <u>II</u> LVFMVFGF <u>II</u> GRRHCGYSKA..DQEGDEELYFHC...	
ensxetp00000040601	<u>III</u> AVVAVAGT <u>II</u> LVFMVFGF <u>II</u> GRRHCGYSKA..DQEGDEELYFHF...	
	AYI.MGSQSSLLLCLQIALQLLINSSSLYYTAALCDLNYNKSLKMHFPSPG	
ensxetp00000040601_corrected	601	650
np_990414	TKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVC	
ensxetp00000040601	..KFPGTKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVC	
	LVKFPGTKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVC	
ensxetp00000040601_corrected	651	700
np_990414	SGRLKLPGKRDPVAIAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLE	
ensxetp00000040601	SGRLKLPGKRDPVAVAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLE	
	SGRLKLPGKRDPVAIAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLE	
ensxetp00000040601_corrected	701	750
np_990414	GVVTRGKPVMIIVIEFMENGALDAFLRKLDGQFTVIQLVGMLRGIAAGMRY	
ensxetp00000040601	GVVTRGKPVMIIVIEYMENGALDAFLRKHDGQFTVIQLVGMLRGIAAGMRY	
	GVVTRGKPVMIIVIEFMENGALDAFLRKLDGQFTVIQLVGMLRGIAAGMRY	
ensxetp00000040601_corrected	751	800
np_990414	LADMGYVHRDLAARNILVNSNLVCKVSDFGLSRIIEDDPDAVYTTTQGGK	
ensxetp00000040601	LADMGYVHRDLAARNILVNSNLVCKVSDFGLSRVIEDDPEAVYTTT.GGK	
	LADMGYVHRDLAARNILVNSNLVCKVSDFGLSRIIEDDPDAVYTTTQGGK	
ensxetp00000040601_corrected	801	850
np_990414	IPVRWTAPEAIQYRKFTSASDVWSYGIVMWEVMSYGERPYWDMNQDVIK	
ensxetp00000040601	IPVRWTAPEAIQYRKFTSASDVWSYGIVMWEVMSYGERPYWDMNQDVIK	
	IPVRWTAPEAIQYRKFTSASDVWSYGIVMWEVMSYGERPYWDMNQDVIK	

Figure 6

Error detected by MisPred routine for Conflict 2. ENSXETP00000040601 of *Xenopus tropicalis* corresponds to the frog ortholog of Ephrin receptor A7, but lacks a typical transmembrane helix between its extracellular FN3 and cytoplasmic Pkinase domains. The mispredicted sequence was corrected by identifying the missing transmembrane sequence using frog EST sequences such as EL820950 [GenBank:EL820950]. The alignment shows the regions containing the transmembrane helices of *Gallus gallus* Ephrin receptor A7 [RefSeq:NP_990414], ENSXETP00000040601 and ENSXETP00000040601_corrected. The predicted transmembrane helices of NP_990414 and ENSXETP00000040601_corrected are in red and underlined, the mispredicted region of ENSXETP00000040601 is in italics.

are non-viable. It is important to point out that such chimeric transcripts of neuronal pentraxin are very rare, or absent, as revealed by blast searches of EST databases. The fact that the rate of such errors is highest in the case of *Fugu rubripes* may be partly due to the short intergenic distance in the compact pufferfish genome.

MisPred routine for Conflict 4 detected a large number of erroneous proteins both among Ensembl entries (1.74–6.95%) and among GNOMON-predicted entries (1.78–15.63%) containing members of Pfam-A domain families suitable for the study of domain integrity. The relatively high rate of erroneous insertion or omission of exons encoding parts of domains indicates that misprediction of exons encoding Pfam-A domains is quite general.

No erroneous human Ensembl protein was detected by the routine for Conflict 5, i.e. none of them were chimeras of genes located on different chromosomes. This is not surprising in view of the fact that chromosomal assembly of human genomic contigs is reliable therefore Ensembl is exempt from the error of trans-chromosomal prediction of human genes. MisPred analysis of the *Homo sapiens* GNOMON-predicted entries has identified only one sequence, XP_001128605 [RefSeq:XP_001128605], as a chimera of genomic regions located on chromosomes 2 and 7. Note that in the NCBI database XP_001128605 has been replaced recently by the nonchimeric NP_001035225 [RefSeq: NP_001035225], encoded on chromosome 7. In contrast with this, several zebrafish entries were identified as suspicious by the MisPred routine for Conflict 5, an observation best explained by errors in the chromosomal assignment and assembly of zebrafish contigs. For example, ENSDARP00000056920 [Ensembl: ENSDARP00000056920] aligns over its entire length with mammalian Casein kinase I isoform gamma-2 proteins, but it is encoded by contigs assigned to chromosomes 2 and 8. XP_001345102 [RefSeq: XP_001345102] corresponds to a fragment of the zebrafish ortholog of ephrin receptor EPHA3, it aligns over its entire length with these receptors, but it is encoded by contigs assigned to chromosome 12 and 25. Incorrect assembly of contigs may lead to the *in silico* fusion of regions located on different chromosomes. For example, the major part of ENSDARP00000077525 [Ensembl: ENSDARP00000077525], encoded on chromosome 7, is homologous with mammalian Solute carrier organic anion transporter family member 3A1 proteins, but an unrelated N-terminal extension of ENSDARP00000077525 is encoded on chromosome 11. XP_001345729 [RefSeq: XP_001345729], a zebrafish protein annotated as similar to TRAF interacting protein is a chimera of traf-interacting protein and plasminogen related growth factor receptor 3.

In summary, GNOMON-predicted sequences and Ensembl sequences are quite similar inasmuch as similarly high proportion of suspicious proteins can be detected by Conflict 1 and Conflict 4 and fewer errors are detected by Conflict 2, Conflict 3 and Conflict 5 (see Tables 3 and 4). Despite these similarities, there are some differences. For example, in the case of Conflict 2 several Ensembl sequences were erroneous because the regions corresponding to their transmembrane helices were mispredicted, whereas this type of error was not found among the GNOMON-predicted sequences. In principle, such differences may reflect differences in the performance of the two gene prediction pipelines or differences in the gene populations covered by the two databases. It should also be pointed out that Ensembl is a comprehensive source of known genes and genes predicted with GeneWise [31], as well as the corresponding transcripts and proteins whereas in the NCBI database transcripts/proteins predicted by GNOMON are distinguished from those of known transcripts/proteins by unique (XM_ or XP_) identifiers.

To permit a more direct comparison of the performance of the two gene prediction pipelines we have compared the results of MisPred analyses only for those protein-coding genes for which both Ensembl and GNOMON have at least one prediction (see Table 5). Comparison of the two datasets confirmed that the two gene prediction pipelines are similar inasmuch as they suffer primarily from errors detectable by MisPred routines for Conflict 1 and Conflict 4, whereas the rates of errors detectable by routines for Conflicts 2, 3 and 5 are very low. Nevertheless, there are minor differences between the Ensembl and NCBI gene prediction pipelines: Ensembl is more likely to fail in the identification of exons encoding transmembrane helices, whereas NCBI's GNOMON appears to be more prone to fuse tandem genes *in silico* (for details of these analyses [see Additional file 1]).

MisPred analysis of GENCODE sequences

The ENCODE (ENCyclopedia Of DNA Elements) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the project is focused on specified 30 megabases (approximately 1%) of the human genome sequence [32]. GENCODE is a sub-project of ENCODE; its overall goal is to identify all protein-coding genes in the regions of the human genome selected within the ENCODE project.

MisPred analysis of the 1097 GENCODE peptides with the routine for Conflict 1 identified one peptide [GENCODE:AC110015.1-002] as containing an extracellular (Cadherin) domain but lacking signal peptide and transmembrane segments. Blast searches revealed that AC110015.1-002 corresponds to the N-terminal part of

Table 5: MisPred analysis of human genes predicted by the EnsEMBL and NCBI's GNOMON pipelines

EnsEMBL									
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	2772	147	5.3%	23	15.65%	ND	ND	ND	ND
Monodelphis domestica	10519	680	6.46%	137	20.15%	ND	ND	ND	ND
Gallus gallus	6139	345	5.62%	113	32.75%	ND	ND	ND	ND
Danio rerio	10289	860	8.36%	317	36.86%	ND	ND	ND	ND
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	2772	1	0.04%	0	0.00%	0	0.00%	0	0.00%
Monodelphis domestica	10519	10	0.1%	0	0.00%	0	0.00%	0	0.00%
Gallus gallus	6139	2	0.03%	0	0.00%	0	0.00%	0	0.00%
Danio rerio	10289	20	0.19%	5	25%	4	20%	1	5%
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	2772			0	0.00%	0	0.00%	0	0.00%
Monodelphis domestica	10519			0	0.00%	0	0.00%	0	0.00%
Gallus gallus	6139			0	0.00%	0	0.00%	0	0.00%
Danio rerio	10289			0	0.00%	0	0.00%	0	0.00%
Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	2772	722	26.05%	48	6.65%	ND	ND	ND	ND
Monodelphis domestica	10519	3726	35.42%	119	3.19%	ND	ND	ND	ND
Gallus gallus	6139	1640	26.72%	159	9.70%	ND	ND	ND	ND
Danio rerio	10289	2565	24.93%	197	7.68%	ND	ND	ND	ND
Conflict 5	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	2772			0	0.00%	0	0.00%	0	0.00%
Danio rerio	10289			0	0.00%	0	0.00%	0	0.00%

Table 5: MisPred analysis of human genes predicted by the EnSEMBL and NCBI's GNOMON pipelines (Continued)

NCBI/GNOMON									
Conflict 1	Number of proteins	Identified as containing an extracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	3012	139	4.61%	32	23.02%	ND	ND	ND	ND
Monodelphis domestica	9703	642	6.62%	112	17.45%	ND	ND	ND	ND
Gallus gallus	5604	310	5.53%	88	28.39%	ND	ND	ND	ND
Danio rerio	8905	742	8.33%	158	21.29%	ND	ND	ND	ND
Conflict 2	Number of proteins	Identified as containing an extra- and an intracellular domain	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	3012	2	0.07%	0	0%	0	0.00%	0	0.00%
Monodelphis domestica	9703	17	0.18%	4	23.53%	2	11.76%	2	11.76%
Gallus gallus	5604	3	0.05%	1	33.33%	1	33.33%	0	0.00%
Danio rerio	8905	16	0.18%	6	37.5%	4	25%	2	12.5%
Conflict 3	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	3012			0	0.00%	0	0.00%	0	0.00%
Monodelphis domestica	9703			1	0.01%	0	0.00%	1	0.01%
Gallus gallus	5604			0	0.00%	0	0.00%	0	0.00%
Danio rerio	8905			2	0.02%	1	0.01%	1	0.01%
Conflict 4	Number of proteins	Proteins containing domains suitable for the study of domain integrity	Percentage	Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	3012	792	26.3%	41	5.18%	ND	ND	ND	ND
Monodelphis domestica	9703	3420	35.25%	39	1.14%	ND	ND	ND	ND
Gallus gallus	5604	1500	26.77%	208	13.87%	ND	ND	ND	ND
Danio rerio	8905	2059	23.12%	300	14.57%	ND	ND	ND	ND
Conflict 5	Number of proteins			Identified as suspicious by MisPred	Percentage*	False Positives	Percentage*	True errors	Percentage*
Homo sapiens	3012			1	0.03%	0	0.00%	1	0.03%
Danio rerio	8905			5	0.06%	5	0.06%	0	0.00%

*Values for suspicious, false positive and true positive sequences are expressed as percentage of the proteins relevant for the given conflict.
 ND – not determined.

The data refer to human genes for which both gene prediction pipelines generated at least one gene model.

CADH2_HUMAN [Swiss-Prot:P19022] but, due to alternative splicing, it lacks the N-terminal secretory signal peptide. It seems unlikely that this transcript encodes a viable protein since – in the absence of a secretory signal peptide – it may not be transported into the extracytoplasmic space.

MisPred analysis of the GENCODE peptides for Conflict 4 identified 67 (6.1% of the total) as containing an abnormally short or abnormally long Pfam-A domain. In the majority of cases the deviant domains are N-terminally or C-terminally truncated simply as a consequence of the incompleteness of the transcripts. Nevertheless, we identified several cases where the domain deviates from normal size as a result of alternative splicing [33].

Examples include AC015691.9-002 [GENCODE:AC015691.9-002] (corresponding to TRIM6_HUMAN [Swiss-Prot:Q9C030]), RP11-247A12.5-001 [GENCODE:RP11-247A12.5-001] (corresponding to CACP_HUMAN [Swiss-Prot:P43155]), XX-FW83563B9.3-006 [GENCODE:XX-FW83563B9.3-006] (corresponding to TAZ_HUMAN [Swiss-Prot:Q16635]), AP006216.3-003 [GENCODE:AP006216.3-003] (corresponding to ZPR1_HUMAN [Swiss-Prot:Q75312]), RP11-298J23.1-003 [GENCODE:RP11-298J23.1-003] (corresponding to PEPC_HUMAN [Swiss-Prot:P20142]) and RP11-247A12.4-008 [GENCODE:RP11-247A12.4-008] (corresponding to PTPA_HUMAN [Swiss-Prot:Q15257]) containing domains that deviate from normal size as a result of alternative splicing. Structural and functional analyses of the putative proteins encoded by these alternatively spliced transcripts suggest that, in many cases, the deviation from normal domain-size may not be compatible with the viability of these proteins, suggesting that the transcripts arose through aberrant splicing [33].

MisPred analysis of the GENCODE peptides for Conflicts 2, 3 and 5 did not identify suspicious sequences.

Conclusion

MisPred may identify proteins as suspicious for three different reasons.

MisPred may identify normal, viable proteins as suspicious due to errors of genomic data and limitations of the bioinformatic tools incorporated in MisPred

MisPred analyses of predicted zebrafish sequences illustrate the point that a 'correct' protein sequence may be suspected to be a chimera encoded by two or more genes located on different chromosomes if the assembly of the contigs is incorrect.

Errors in the bioinformatic identification of sequence features used to detect the various types of Conflicts may

raise unjustified doubts about the viability of some protein sequences. For example, this type of error is encountered in the case of some secreted or transmembrane proteins (identified as such by the presence of obligatory extracellular domains) whose atypical signal peptides or transmembrane helices are not detected with high confidence by the signal peptide or transmembrane helix prediction programs incorporated into the MisPred routines. Similarly, the hmmpfam program may on occasion identify marker domains erroneously, leading to the incorrect prediction of the subcellular localization of a protein domain.

Analyses of the benchmark Swiss-Prot proteins, however, have revealed that the false positive rates of the MisPred routines are lower than 0.001%. For details of the specificity analyses see Materials and methods.

MisPred may identify normal, viable proteins as suspicious due to limitations of the dogmas on which MisPred routines are based

A survey of the results of MisPred analyses has revealed that there are some exceptions to the dogmas on which the MisPred approach is based. For example, some secreted proteins may truly lack secretory signal peptides since they are subject to leaderless protein secretion [22], some predominantly extracellular, cytoplasmic or nuclear Pfam-A domain families are not always restricted to a single subcellular location and thus may be multilocal etc. Similarly, it cannot be excluded at present that transchromosomal chimeras can be formed and may have normal physiological functions [16]. Nevertheless, the fact that MisPred analyses of protein sequences of the Swiss-Prot database identified very few such exceptions indicates that the dogmas of MisPred are generally valid.

MisPred may identify truly abnormal, nonviable proteins

MisPred analyses of Swiss-Prot, TrEMBL, EnsEMBL, NCBI and GENCODE sequences identified numerous hypothetical protein sequences that are likely to be nonviable since they violate some of the basic dogmas about viable proteins.

For example, analysis of the TrEMBL sequences revealed that incomplete or abnormal hypothetical proteins translated *in silico* from incomplete cDNAs or aberrant transcripts are quite abundant in this database. A recent study [33] has also revealed that alternative splicing frequently generates transcripts that encode nonviable proteins due to violation of domain integrity, loss of signal peptides, etc.

Interestingly, there are also numerous chimeric entries in the TrEMBL database, a large proportion of which are derived from normal tissues. There are several mecha-

nisms for the creation of chimeric proteins. Unequal crossing over resulting in the fusion of parts of tandem genes and chromosomal translocation resulting in the fusion of genes located on different chromosomes are the two best documented mechanisms for the creation of chimeric genes and proteins. Recent studies have convincingly shown that transcription of tandem genes into a single RNA sequence and translation of the chimeric mRNA into a chimeric protein [34,35] is also a major mechanism for the creation of chimeric proteins. Unneberg and Claverie [36] have recently suggested that chimeric proteins might also be formed through transchromosomal transcription, i.e. if genes located on different chromosomes are expressed in the same "transcription factory" they may give rise to chimeric transcripts.

It should be pointed out that the relatively high proportion of incomplete and abnormal entries in TrEMBL also has an impact on gene prediction. Since most gene prediction pipelines rely on extrinsic information [5] such as those provided by TrEMBL, some of the errors of TrEMBL may be inherited by the databases of predicted genes. It is noteworthy in this respect that the majority of errors of Ensembl and GNOMON-predicted sequences can be detected with the use of Conflict 1 and Conflict 4, probably reflecting the fact the TrEMBL sequences also suffer from the same types of errors. The relatively high proportion of transchromosomal chimeric sequences among TrEMBL entries, however, does not have a major impact on gene prediction, provided that contig assembly is unambiguous.

Sequences not detected by MisPred do not necessarily correspond to normal, viable proteins

Although MisPred identifies many suspicious sequences, it should be emphasized that the routines detect only a fraction of the truly erroneous sequences. First, the MisPred routines described in this manuscript exploit only five of the various dogmas about viable proteins. Second, only a fraction of proteins contains members of well-characterized Pfam-A domain families, i.e. families on which the routines for Conflicts 1, 2, 3 and 4 rely. Third, the number of suspicious sequences identified by Conflicts 1, 2 or 3 significantly underestimates the actual number of sequences that may be affected by these types of errors since we used only validated, obligatory extracellular, cytoplasmic and nuclear Pfam-A domain families to predict the subcellular localization of proteins and did not include numerous extracellular, cytoplasmic or nuclear domain families the members of which may also occur in other subcellular compartments. Fourth, in the case of Conflict 4 MisPred uses a high cut-off value: a domain is judged to be nonviable only if its Pfam-A domain deviates from normal size by at least 40% in length. For details of the sensitivity analyses see Materials and methods.

Summary

Recent studies have shown that a significant proportion of eukaryotic genes may be mispredicted at the transcript level [5,6]. Since the MisPred routines described here are able to detect many of these errors and may aid the correction of these errors we suggest that the MisPred approach may significantly improve the quality of protein sequence data based on gene-predictions. In order to increase the sensitivity of the approach we are currently developing additional routines, based on the violation of other dogmas about proteins. The MisPred approach may also be used as a discovery tool since it can also serve to explore the limitations of the dogmas on which the various MisPred routines are based.

Methods

Protein sequence data analyzed

Protein sequence databases

In the analysis of the Swiss-Prot section of the UniProtKB we have included Metazoan species that have at least 1000 Swiss-Prot entries. The UniProtKB Swiss-Prot [37] entries from UniProtKB Version 9.5 (January, 2007) were downloaded from <http://www.expasy.uniprot.org/database/download.shtml>.

The UniProtKB TrEMBL [37] entries from UniProtKB Version 10.5 (May, 2007) were downloaded from <http://www.expasy.uniprot.org/database/download.shtml>.

The 1097 GENCODE protein sequences were obtained from <http://genome.imim.es/biosapiens/gencode/dataset/v2.2.html>.

The protein sequences of the Ensembl database were downloaded from the Ensembl website [38], release 41 (October, 2006), found at <ftp://ftp.ensembl.org/pub/release-41>.

The NCBI's protein sequences of four species (*Homo sapiens*, 15 September 2006, *Monodelphis domestica*, 06 March 2007, *Gallus gallus*, 30 November 2006, *Danio rerio*, March 2007), were obtained by downloading the relevant protein.fa.gz files from the NCBI Genome Data/Annotation Projects website [39], found at <ftp://ftp.ncbi.nih.gov/genomes/>. In order to analyze only the sequences predicted by GNOMON <http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.html>, an in-house program was used to extract only GNOMON-predicted FASTA sequences with 'XP_' identifiers.

Comparison of the Ensembl and NCBI/GNOMON gene predictions

In order to compare the reliability of the two gene prediction pipelines, we analyzed the proportion of suspicious sequences among predicted proteins represented in both Ensembl and NCBI's GNOMON-predicted section from

four evolutionarily distant species (*Homo sapiens*, *Mono delphis domestica*, *Gallus gallus* and *Danio rerio*).

To identify the number of genes in each species for which both GNOMON and EnsEMBL have at least one predicted protein sequence, blastp [40] searches were performed on the GNOMON-predicted sequences using EnsEMBL sequences as queries. Protein sequences that displayed 100% identity over at least 25 amino acid residue-long ungapped segments were considered to be encoded by the same gene. The blastp standalone program, version 2.2.13, was obtained from: <ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.13>.

Prediction of the subcellular localization of proteins by the presence of obligatory extracellular, cytoplasmic or nuclear protein-domains

Recent studies have revealed that there is a strong correlation between the domain-composition of proteins and their subcellular location: some domains are restricted to proteins targeted to the extracellular space, others occur only in proteins present in the cytoplasmic space, whilst others are restricted to proteins of the nucleus [17,18]. Transmembrane multidomain proteins are special in the sense that obligatory extracellular and cytoplasmic domains can legitimately co-occur in a single protein. Accordingly, the presence of obligatory extracellular, cytoplasmic or nuclear domains in a protein may be used to predict its subcellular localization, independent of the detection of sorting signals.

Since domains are most likely to co-occur in multidomain proteins if they belong to the same localization-category, analysis of domain co-occurrence networks is useful for the systematic assignment of domains to different subcellular compartments [17,18]. Our domain co-occurrence analyses of Metazoan UniProtKB entries have identified 166 obligatory extracellular, 115 obligatory cytoplasmic and 126 obligatory nuclear Pfam-A domain families as being restricted to the respective subcellular compartment, the majority of which are also identified as such in the SMART database <http://smart.embl-heidelberg.de/> [see Additional files 3, 4 and 5], respectively. In the Mis-Pred analyses described in the present work only these obligatory extracellular or cytoplasmic or nuclear domain families were used to predict subcellular localization. Pfam-A domains that are known not to be restricted to a particular cellular compartment, such as immunoglobulin domains, fibronectin type III domains, von Willebrand factor type A domains (i.e. domains that are 'multilocal'), are not reliable predictors of subcellular localization and thus they were not utilized in these analyses.

The programs of the HMMER 2.3.2 software package were used to detect obligatory extracellular, cytoplasmic and

nuclear Pfam-A domains. HMM databases of Pfam-A domains were created by retrieving the HMMs of the domains from the Pfam (Pfam_ls) HMM and the Pfam fragment (Pfam_fs) HMM libraries [41]. The presence of Pfam-A domains in protein sequences was detected by searching the HMM databases against protein sequences using the hmmpfam program using 0.00001 as per-domain E-value threshold. We filtered the results for overlapping domain matches and the match with the lowest E-value was accepted.

The Pfam HMM libraries (Release 20.0) were obtained from <ftp://ftp.sanger.ac.uk/pub/databases/Pfam/>. The HMMER software package was obtained from <http://hmm.janelia.org/>.

Detection of suspicious (incomplete, abnormal and mispredicted) proteins

Conflict 1: Conflict between the predicted subcellular localization of proteins and the absence of the corresponding sequence signals

Proteins containing obligatory extracellular Pfam-A domains were analyzed by the PrediSi program [42] to identify the presence of eukaryotic signal peptide sequences (using 0.3 as threshold) and by the TMHMM program [43] to detect the presence of transmembrane helices. Protein sequences containing obligatory extracellular domains but neither a signal peptide nor a transmembrane helix were identified as suspicious. The PrediSi program was downloaded from <http://www.predisi.de>. The TMHMM program was obtained from <http://www.cbs.dtu.dk/services/TMHMM>.

Conflict 2: Presence of obligatory extracellular and cytoplasmic domains and the absence of transmembrane helices

Proteins containing both obligatory extracellular and obligatory cytoplasmic Pfam-A domains were analyzed by the TMHMM program to detect transmembrane helices. Sequences, which contain both extra- and cytoplasmic domains (i.e. putative transmembrane proteins), but do not have a transmembrane helix were identified as suspicious.

Conflict 3: Co-occurrence of extracellular and nuclear domains in a single protein

Protein sequences containing both obligatory extracellular and obligatory nuclear domains were identified as suspicious.

Conflict 4: Violation of domain integrity

In order to identify the Pfam-A domain families that have a well-defined, conserved sequence length range, we selected only those families whose members do not deviate from the average size by more than 2 standard deviation (SD) values in the high quality Swiss-Prot database (version 48.9). Based on these criteria, about 90% of the Pfam-A domain families present in Swiss-Prot proteins

proved to be suitable for the study of domain integrity. The list of the Pfam-A families is deposited in Additional file 6 [see Additional file 6]. We created databases of human, vertebrate and metazoa+fungi Swiss-Prot domain sequences belonging to the reliable Pfam-A families and ran a blastp search with the current set of proteins as queries against the appropriate reliable Swiss-Prot domain sequences. We selected those partial domain matches which share over 60% identity with the query sequence, with an E-value $< 1e^{-5}$, and differ by at least 40% in length. The protein sequences containing these domains with deviant lengths were identified as suspicious. For the details of this method [see Additional file 1].

Conflict 5: Chimeric proteins encoded by two or more different genes located on different chromosomes

The protein sequences were matched to the genome of the given species using the BLAT program [44]. We selected matches with $>95\%$ identity over ≥ 15 amino acid residue in length and in the case of overlapping matches (if the overlap was >5 residues) we selected the longest match. To eliminate problems encountered with genes encoded by the mitochondrial genome [see Additional file 1] we used an additional BLAT search and discarded those entries which gave $>90\%$ match with the mitochondrial genome over more than 90% of their length. Proteins were considered suspicious if two or more of their segments were encoded on different chromosomes. The BLAT program was obtained from <http://genome-test.cse.ucsc.edu/~kent/exe/>. The following genome assemblies were used in these analyses: *Homo sapiens* NCBI 36, Mar 2006; *Mus musculus* NCBI 36, Febr 2006; *Rattus rattus* RGSC 3.4, Nov 2004; *Gallus gallus* WASHUC2, May 2006; *Danio rerio* Zv6, Mar 2006;; *Caenorhabditis elegans* WS170, Jan 2007; *Drosophila melanogaster* BDGP 5, Apr 2006.

Testing the specificity and sensitivity of MisPred methods
Specificity

To calculate the false positive rate (α) and specificity ($1-\alpha$) of MisPred from the equation $\alpha = FP/(FP+TN)$, we have determined the number of false positives (FP) and true negatives (TN) from the results obtained by application of MisPred to Swiss-Prot entries. In these calculations FP equals the number of entries that were identified with the given method as suspicious, although they do not to violate the given dogma. Considering that Swiss-Prot is a very clean database, the entries not identified by MisPred as suspicious were assumed to be true negatives (TN), i.e. they do not violate the given dogma. Specificity was also calculated by analyzing datasets obtained by mixing correct Swiss-Prot entries and erroneous sequences generated from Swiss-Prot sequences (see below). The false positive rates of the MisPred routines were calculated to be ≤ 0.001 , i.e. their specificity is very high (≥ 0.999).

Sensitivity

Since at the protein level the major types of errors of gene prediction (failure to find a true exon, erroneous inclusion of a false exon, misprediction of an exon, fusion of exons of tandem genes etc.) are manifested as internal or terminal deletions, internal insertions, terminal extensions and fusions, we have generated datasets of sequences from human Swiss-Prot entries to mimic these errors.

To test the effect of terminal deletions, a group of datasets was created through deletion of 50, 100, 150, 200 etc. residues from their N-terminal end or their C-terminal end. Another group of datasets were obtained by deleting the second, third, fourth etc. 50 or 100 residue-segments of the proteins to study the effect of internal deletions. Terminal extensions or internal insertions were mimicked by addition/insertion of 50 or 100 amino acid segments (with random sequences and average amino acid composition) to the N-terminal and C-terminal end or after positions 50, 100, 150 etc. of these proteins. To mimic the effect of fusions datasets were generated by fusing 5000 randomly selected entries to a different set of 5000 randomly selected proteins.

Subsets of the above datasets were created to gain insight into the factors that influence the sensitivity of the different MisPred methods. The false negative rate (β) and sensitivity ($1-\beta$) of MisPred were calculated from the equation $\beta = FN/(TP + FN)$. Since all entries in these datasets are erroneous (i.e. they differ from the correct sequence), entries detected by MisPred as suspicious are true positives (TP), whereas those not detected by MisPred are false negatives (FN).

The MisPred routine for Conflict 1 detected only 4.5% of the sequences from which the N-terminal 50 residues (that might contain secretory signal peptide or signal-anchor sequences) were removed. One major source of such a low sensitivity is that only a fraction of proteins are secreted or type II transmembrane proteins, the integrity of which can be tested with this method. When we analyzed a dataset generated by N-terminal truncation of proteins containing signal peptide or signal anchor sequences but lacking transmembrane segments, sensitivity increased to 33.1%. The reason why only about a third of the erroneous proteins are detected by MisPred is that the majority of these proteins does not contain an obligatory extracellular Pfam-A domain and is thus 'invisible' to this method. When we restricted MisPred analysis to erroneous entries generated from secreted and type II transmembrane proteins containing an obligatory extracellular Pfam-A domain, sensitivity increased to 87.2%.

Only a very small fraction (0.79%) of the proteins possessing transmembrane helices are detected as erroneous by MisPred routine for Conflict 2 after removing their transmembrane helices. This is due to the fact that few of the transmembrane proteins contain both an obligatory extracellular and an obligatory cytoplasmic Pfam-A domain. If, however, we applied this method to erroneous entries generated from transmembrane proteins containing both an obligatory extracellular and an obligatory cytoplasmic Pfam-A domain, sensitivity increased to 83.7%.

The sensitivity of the MisPred routine for Conflict 3 was found to be 1.06% when we applied it to chimeric proteins generated by random fusion of proteins. Such a low sensitivity is due to several factors. First, nuclear proteins account for only ~14% of the entire proteome [45]. Second, our analyses have shown that only 26.7% of the nuclear proteins identified by Fink et al. [45] (2008) contain obligatory nuclear domains [see Additional file 5]. Third, only a fraction of proteins contains an obligatory extracellular Pfam-A domain. In harmony with this explanation, the sensitivity of the MisPred routine for Conflict 3 was found to be 99.9% when we applied it to chimeric proteins generated by fusion of human Swiss-Prot proteins containing obligatory extracellular domains to proteins containing obligatory nuclear domains.

The sensitivity of the MisPred routine for Conflict 4 was found to depend on the extent of terminal truncation of the protein sequences. Progressive truncation from the C-terminal or N-terminal end increased the sensitivity to a maximum of ~4% when ~250 residues were deleted, deletion of longer segments did not further increase sensitivity. The explanation for this observation is that this type of error becomes undetectable if the entire domain or the major part of the domain is removed. Sensitivity was found to be ~0.6% (or ~3%) in the case of internal deletions of 50 residue (or 100 residue) segments, and ~0.1% (or 4%) in the case of internal insertions of 50 (or 100) residue-long segments with random amino acid sequences, irrespective of the positions of the moving window deletions or the insertions. Addition of such random sequences to the N-terminal end, C-terminal end of human Swiss-Prot entries did not generate errors detectable by MisPred routine for Conflict 4. The relatively low sensitivity of the MisPred routine for Conflict 4 is due to the fact that only a fraction of human Swiss-Prot proteins contain a Pfam-A domain suitable for the detection of domain size deviation (see Table 1). Another factor that contributes to the low sensitivity of this approach is that a protein is detected as erroneous only if its Pfam-A domain deviates from normal size by at least 40% in length (see above).

The sensitivity of the MisPred routine for Conflict 5 was found to be very high (91.5%), when tested on artificial chimeras generated by random fusion of human Swiss-Prot proteins. Analysis of the few false negatives revealed that some of the chimeric proteins were not detected because the constituent proteins are encoded on the same chromosome, in other cases the BLAT match was below the 95% threshold. If we restricted the analyses to chimeras generated by fusion of genes encoded on different chromosomes, sensitivity increased to 92.9%.

Resolution of Conflicts

In order to test whether a suspicious protein sequence identified by one of the MisPred routines is truly erroneous (or a false positive) we subjected these sequences to additional analyses. Such analyses were performed for all suspicious Swiss-Prot entries identified by all five MisPred routines as well as for all Ensembl, GNOMON-predicted and human TrEMBL entries identified by MisPred routines for Conflicts 2, 3 and 5 (for details [see Additional file 1]).

Sequences identified by MisPred routines for Conflict 1 and Conflict 2 as suspicious were analyzed by the SignalP [46] <http://www.cbs.dtu.dk/services/SignalP/> and Phobius [47] <http://phobius.cgb.ki.se/> programs to detect signal peptides and/or transmembrane helices potentially missed by the PrediSi and/or the TMHMM programs.

To identify cases of non-classical, i.e. not signal peptide triggered protein secretion we used the SecretomeP 1.0b Server [22] <http://www.cbs.dtu.dk/services/SecretomeP-1.0/>. The SecretomeP 1.0b server produces *ab initio* predictions of non-classical protein secretion in eukaryotes. It should be noted, however, that this approach is not necessarily able to decide whether a given entry is an example of leaderless secretion or it is a fragment of a classically secreted protein (that lacks its signal peptide). To exclude the latter possibility, we have also queried EST databases to decide whether the initiating methionine is properly defined or the open reading frame may be extended in the upstream direction.

To further test whether a suspicious protein sequence (sequence A) identified by one of the MisPred routines is truly erroneous it was used as query to search protein and nucleic acid sequence databases with blastp and tblastn, respectively, to identify homologous sequences from the same as well as from other species. If the search yielded a perfect match (100% identity over at least 25 amino acid residues) with a different protein sequence from the same species (sequence B), but the latter protein was judged 'normal' by the same MisPred routine that identified sequence A as suspicious then we concluded that the error has been both validated and corrected (by sequence B). If

the sequence similarity search identified close homologs of sequence A (paralogs from the same species, orthologs and/or paralogs from related species) but the latter protein(s) were judged 'normal' by the same MisPred routine that identified sequence A as suspicious then we concluded that the error has been validated. In the majority of such cases the erroneous sequence A could be corrected through targeted search of the appropriate genomic region of the relevant species with various gene prediction programs as well as search of EST databases, using the 'normal' homologous sequence(s) as queries. The protocol used for the correction of errors will be described in another publication (manuscript in preparation).

Authors' contributions

AN, KF, HT and EK have developed the MisPred methods for Conflicts 1, 2, 3 and 5, HH developed the MisPred method for Conflict 4. LB was involved in bioinformatic analyses of the protein sequences identified as suspicious by the different MisPred routines and in correction of erroneous sequences. LP was involved in conceiving and planning the project.

Additional material

Additional file 1

Comments on entries identified by MisPred routines as suspicious and detailed description of Conflict 4. The file describes detailed analyses and comments of the entries identified as suspicious by the different MisPred routines, and contains the detailed description of the procedure of the identification of suspicious sequences by Conflict 4.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S1.pdf>]

Additional file 2

List of erroneous Swiss-Prot sequences identified by MisPred. The file contains the list of erroneous Swiss-Prot sequences identified by MisPred.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S2.pdf>]

Additional file 3

List of extracellular Pfam-A domain families. The file contains the list of extracellular Pfam-A domain families.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S3.pdf>]

Additional file 4

List of intracellular Pfam-A signaling domain families. The file contains the list of intracellular Pfam-A signaling domain families.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S4.pdf>]

Additional file 5

List of nuclear Pfam-A domain families. The file contains the list of nuclear Pfam-A domain families.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S5.pdf>]

Additional file 6

List of Pfam-A domain families suitable for the study of domain integrity. The file contains the list of Pfam-A domain families suitable for the study of domain integrity.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-353-S6.pdf>]

Acknowledgements

This work was carried out as part of the BioSapiens project. The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHS-G-CT-2003-503265. The authors thank the partial support of the National Office for Research and Technology under grant no.: RET14/2005.

References

- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramsier J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S,

- Chen YJ, International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huseon DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabriellian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkuch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hattori T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
3. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
4. Pennisi E: **Working the (gene count) numbers: finally, a firm answer?** *Science* 2007, **316**(5828):1113.
5. Guigó R, Flicek P, Abril JF, Raymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG: **EGASP: the human ENCODE genome annotation assessment project.** *Genome Biol* 2006, **7**(Suppl 1):S2.
6. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nat Rev Genet* 2008, **9**:62-73.
7. Wang T, Waters CT, Rothman AM, Jakins TJ, Romisch K, Trump D: **Intracellular retention of mutant retinoschisin is the pathological mechanism underlying X-linked retinoschisis.** *Hum Mol Genet* 2002, **11**:13097-105.
8. Ohnishi A, Emi Y: **Rapid proteasomal degradation of translocation-deficient UDP-glucuronosyltransferase 1A1 proteins in patients with Crigler-Najjar type II.** *Biochem Biophys Res Commun* 2003, **310**:735-41.
9. Saarela J, von Schantz C, Peltonen L, Jalanko A: **A novel aspartylglucosaminuria mutation affects translocation of aspartylglucosaminidase.** *Hum Mutat* 2004, **24**:350-1.
10. Jayakumar A, Kang Y, Henderson Y, Mitsudo K, Liu X, Briggs K, Wang M, Frederick MJ, El-Naggar AK, Bebek Z, Clayman GL: **Consequences of C-terminal domains and N-terminal signal peptide deletions on LEKTI secretion, stability, and subcellular distribution.** *Arch Biochem Biophys* 2005, **435**:89-102.
11. Hansen L, Yao W, Eiberg H, Funding M, Riise R, Kjaer KW, Hejtmancik JF, Rosenberg T: **The congenital "ant-egg" cataract phenotype is caused by a missense mutation in connexin46.** *Mol Vis* 2006, **12**:1033-9.
12. Mukherjee O, Pastor P, Cairns NJ, Chakraverty S, Kauwe JS, Shears S, Behrens MI, Budde J, Hinrichs AL, Norton J, Levitch D, Taylor-Reinwald L, Gitcho M, Tu PH, Tenenholz Grinberg L, Liscic RM, Armentariz J, Morris JC, Goate AM: **HDDD2 is a familial frontotemporal lobar degeneration with ubiquitin-positive, tau-negative inclusions caused by a missense mutation in the signal peptide of progranulin.** *Ann Neurol* 2006, **60**:314-22.
13. Favre B, Plantard L, Aeschbach L, Brack N, Christen-Zaech S, de Viragh PA, Sergeant A, Huber M, Hohl D: **SLURPI is a late marker of epidermal differentiation and is absent in Mal de Meleda.** *J Invest Dermatol* 2007, **127**:301-8.
14. Hudziak RM, Ullrich A: **Cell transformation potential of a HER2 transmembrane domain deletion mutant retained in the endoplasmic reticulum.** *J Biol Chem* 1991, **266**:24109-15.
15. Brenner C, Fuller RS: **Structural and enzymatic characterization of a purified prohormone-processing enzyme: secreted, soluble Kex2 protease.** *Proc Natl Acad Sci USA* 1992, **89**:922-6.
16. Goldberg AL: **Protein degradation and protection against misfolded or damaged proteins.** *Nature* 2003, **426**:895-9.
17. Mott R, Schultz J, Bork P, Ponting CP: **Predicting protein cellular localization using a domain projection method.** *Genome Res* 2002, **12**:1168-74.
18. Tordai H, Nagy A, Farkas K, Banyai L, Patthy L: **Modules, multidomain proteins and organismic complexity.** *FEBS J* 2005, **272**:5064-5078.
19. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries.** *Bioinformatics* 2000, **16**:613-8.
20. Wolf Y, Madej T, Babenko V, Shoemaker B, Panchenko AR: **Long-term trends in evolution of indels in protein sequences.** *BMC Evol Biol* 2007, **7**:19.
21. Watters AL, Deka P, Corrent C, Callender D, Varani G, Sosnick T, Baker D: **The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection.** *Cell* 2007, **128**:613-24.
22. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion.** *Protein Eng Des Sel* 2004, **17**:349-56.
23. Tordai H, Patthy L: **Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides.** *FEBS Lett* 2004, **575**:109-11.
24. Banyai L, Patthy L: **Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues.** *FEBS Lett* 2004, **565**:127-32.
25. C. elegans Sequencing Consortium: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282**:2012-2018.
26. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkuch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabriellian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Paceb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-

- Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287**:2185-2195.
27. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucl Acids Res* 2004, **32**:D142-4.
 28. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoeve F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002, **297**:1301-1310.
 29. Chen B, Bixby JL: **Neuronal pentraxin with chromo domain (NPCD) is a novel class of protein expressed in multiple neuronal domains.** *J Comp Neurol* 2005, **481**:391-402.
 30. Chen B, Bixby JL: **A novel substrate of receptor tyrosine phosphatase PTPRO is required for nerve growth factor-induced process outgrowth.** *J Neurosci* 2005, **25**:880-888.
 31. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14**:988-995.
 32. ENCODE project consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**:636-640.
 33. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PL, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, Lopez G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Stirling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT, Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A: **The implications of alternative splicing in the ENCODE protein complement.** *Proc Natl Acad Sci USA* 2007, **104**:5495-500.
 34. Akiya P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: **Transcription-mediated gene fusion in the human genome.** *Genome Res* 2006, **16**:30-6.
 35. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R: **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 2006, **16**:37-44.
 36. Unneberg P, Claverie JM: **Tentative Mapping of Transcription-Induced Interchromosomal Interaction using Chimeric EST and mRNA Data.** *PLoS ONE* 2007, **2**:e254.
 37. The UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucl Acids Res* 2007, **35**:D193-D197.
 38. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucl Acids Res* 2007, **35**:D610-D617.
 39. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucl Acids Res* 2007, **35**:D5-D12.
 40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25**:3389-3402.
 41. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucl Acids Res* 2006, **34**:D247-51.
 42. Hiller K, Grote A, Scheer M, Munch R, Jahn D: **PrediSi: prediction of signal peptides and their cleavage positions.** *Nucl Acids Res* 2004, **32**:W375-9.
 43. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
 44. Kent WJ: **BLAT - the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656-664.
 45. Fink JL, Karunaratne S, Mittal A, Gardiner DM, Hamilton N, Mahony D, Kai C, Suzuki H, Hayashizaki Y, Teasdale RD: **Towards defining the nuclear proteome.** *Genome Biol* 2008, **9**:R15.
 46. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**:783-795.
 47. Kall L, Krogh A, Sonnhammer EL: **Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server.** *Nucl Acids Res* 2007, **35**:W429-32.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



Additional file 1. Comments on entries identified by MisPred routines as suspicious and detailed description of Conflict 4. Part 1 contains comments on suspicious Swiss-Prot, EnSEMBL- and GNOMON-predicted proteins, and on EnSEMBL and NCBI entries predicted for the same set of genes. Part 2 contains detailed description of the procedure of the identification of suspicious proteins by Conflict 4.

Part 1. Comments on entries identified by MisPred routines as suspicious

Comments on Swiss-Prot entries identified by MisPred routines as suspicious

Conflict 1

Analysis of the *Mus musculus* Swiss-Prot proteins by MisPred identified 12 proteins as suspicious. Four of the suspicious proteins proved to be false positives: although PrediSi failed to detect their signal peptides these were detected with the SignalP and/or Phobius programs (these false positives are annotated in Swiss-Prot as possessing signal peptides).

Three mouse proteins (GAPR1_MOUSE, TENR_MOUSE, VEGFD_MOUSE) containing extracellular domains apparently truly lack classical signal peptides (i.e. none of the signal prediction programs detected a signal sequence) and are probably subject to leaderless protein secretion [1]. Analysis of the remaining five entries revealed that they are truly abnormal and two of these (LY6G_MOUSE, MCPT3_MOUSE) are also annotated as fragments in Swiss-Prot.

C209C_MOUSE lacks both signal peptide and signal anchor sequences, whereas all of the closely related entries are type II transmembrane proteins. A corrected sequence of this protein could be predicted by targeted search of mouse genomic and EST sequences (Figure 2).

NOE2_MOUSE also lacks a signal peptide, whereas the rat ortholog (NP_001015017) and a different isoform of this mouse protein (EDL25126) do possess a signal sequence.

TMPS7_MOUSE lacks a transmembrane helix: it is an N-terminally truncated version of mouse matriptase 3 (AAY66996).

In the case of *Rattus norvegicus* proteins MisPred identified 18 proteins as suspicious. One of these proteins, TENR_RAT, containing extracellular domains apparently truly lacks classical signal peptides (i.e. none of the signal prediction programs detected a signal sequence) and is probably subject to leaderless protein secretion [1].

Fourteen proteins ANGP2_RAT, CADH4_RAT, CATG_RAT, CO5_RAT, CO8B_RAT, COBA1_RAT, FA9_RAT, GDF11_RAT, KLK10_RAT, KLK3_RAT, MUC2L_RAT, NID1_RAT, PDGFB_RAT, SFRP1_RAT proved to be true positives (also annotated as fragments in Swiss-Prot). CSF2_RAT, MIP2A_RAT are also fragments of larger precursors although not annotated as such in Swiss-Prot.

Analysis of *Gallus gallus* Swiss-Prot proteins by MisPred identified 22 as being suspicious. Analysis of these proteins revealed that 19 proteins are in fact abnormal, and 12 of these are also annotated as fragments by Swiss-Prot. The remaining seven entries (SECR_CHICK, PYY_CHICK, IGF2_CHICK, NMU_CHICK, AMP1_CHICK, RNL2_CHICK, IPK1L_CHICK) are also fragments of larger precursors but are not annotated as such by Swiss-Prot.

In the case of *Danio rerio* proteins MisPred identified four proteins as suspicious. One protein, OSTC_DANRE, is truly abnormal and is also annotated as a fragment by Swiss-Prot.

Analysis of *Caenorhabditis elegans* Swiss-Prot proteins by MisPred identified 9 proteins as being suspicious. Detailed analyses of these sequences revealed that LAML1_CAEEL, NAS10_CAEEL, YP95_CAEEL, YOW6_CAEEL, NAS13_CAEEL, CUBN_CAEEL, YL54_CAEEL are true positives: they lack secretory signal peptides and transmembrane helices, although their close homologs contain such sequence signals. (In the majority of these cases the mispredicted sequences could be corrected by targeted search of genomic sequences using close homologs as queries).

NUCG_CAEEL is a mitochondrial protein that truly lacks a eukaryotic-type secretory signal peptide although it contains an extracellular Endonuclease_NS domain, cautioning that this Pfam-A domain family is not obligatory extracellular.

YL15_CAEEL, containing an extracellular Kunitz_BPTI domain, was also identified as suspicious since it lacked signal peptide and transmembrane sequences. Blast searches have confirmed that this is a mispredicted chimeric protein, created through *in silico* joining of two tandem genes located on *Caenorhabditis elegans* chromosome X. The Homeobox-containing region is derived from a gene closely related to the gene encoding the nuclear protein HM07_CAEEL, whereas the Kunitz_BPTI-containing region is derived from the *C. elegans* ortholog of an extracellular protein of *C. briggsae*, Q619J1_CAEER. The structures of the constituent genes could be predicted using the sequences of these homologs (Figure 3). (Note that the co-occurrence of a predominantly nuclear Homeobox domain and an obligatory extracellular Kunitz_BPTI domain also violates the rule underlying Conflict 3).

Analysis of *Drosophila melanogaster* proteins identified 5 entries as suspicious. Two of these (CHIT1_DROME, CHIT3_DROME) are fragments of longer proteins possessing signal peptides. PGPLE_DROME appears to be a false positive that truly lacks a signal peptide, probably because its extracellular Amidase_2 domain is not necessarily associated with secretory signal peptides.

Conflict 4

The MisPred routine for Conflict 4 revealed that of the 15638 human proteins 6973 contained at least one domain belonging to Pfam-A domain families suitable for the study of domain integrity [see Additional file 6]. Ten of these proteins were identified by MisPred as suspicious. Four proteins contained an incomplete domain, whereas the remaining six proteins

were false positives due to failure to detect the full-length Pfam-A domain or a false hit with a Pfam-A domain. Three of the four abnormal proteins (HB2T_HUMAN, POK9_HUMAN, M4A4E_HUMAN) were also annotated as fragments by Swiss-Prot, one protein (VHLL_HUMAN) identified by MisPred as abnormal was not annotated as a fragment in Swiss-Prot. VHLL_HUMAN is closely related to VHL_HUMAN but its C-terminal VHL domain is truncated [2].

Analysis of mouse Swiss-Prot proteins has identified three proteins as suspicious. Only one of these proved to be truly abnormal: MYO1A_MOUSE contains a truncated Myosin_TH1 domain (the protein is also annotated as a fragment by Swiss-Prot).

Analysis of rat Swiss-Prot proteins with the MisPred routine for Conflict 4 has identified 14 proteins that contain a domain of abnormal size. 13 of these proteins are fragments and are also annotated as such by Swiss-Prot. EPHA5_RAT contains a C-terminally truncated SAM_1 domain, although not annotated as a fragment by Swiss-Prot. It is noteworthy that orthologs from mouse, human and chicken contain an intact SAM_1 domain. The sequence could be corrected with targeted search of the rat genome using the sequences of the full-length orthologs (Figure 4).

Analysis of chicken Swiss-Prot proteins has identified 8 proteins as abnormal, all of which were also annotated as fragments by Swiss-Prot.

Only one of the *Danio rerio* Swiss-Prot proteins was identified as abnormal by Conflict 4. DHH_BRARE, containing a truncated HH_signal domain, is a fragment of the full-length protein; it is also annotated as a fragment by Swiss-Prot.

Analyses of *C. elegans* Swiss-Prot proteins have identified two proteins that contain a domain of deviant size. YQS2_CAEEL, containing an N-terminally truncated Patched domain, is actually a fragment of PTC1_CAEEL, although not annotated as a fragment by Swiss-Prot. (The latter protein contains a full-length Patched domain.) PME6_CAEEL, containing a truncated

WGR domain, is a fragment of PME1_CAEEL, the latter protein contains a full-length WGR domain. (Note that PME6_CAEEL in release 45.5 has been deleted.)

None of the *Drosophila melanogaster* Swiss-Prot entries these proteins were identified by the MisPred routine for Conflict 4.

Conflict 5

MisPred identified five human Swiss-Prot entries as suspected to be translated from chimeras of genes located on different chromosomes. Two of these, CR030_HUMAN and MED12_HUMAN, were true chimeras derived from unrelated genes located on different chromosomes. In the case of MED12_HUMAN, the N-terminal 22 residues are encoded by chromosome 1, whereas residues 23-2198 are encoded by chromosome X. (It should be pointed out that the sequence of MED12_HUMAN was determined from a cDNA derived from cervix carcinoma).

In the case of CR030_HUMAN, the N-terminal 1-563 residues of this hypothetical protein are encoded by chromosome 18, whereas residues 564-583 are derived from *Homo sapiens* asparagine-linked glycosylation 12 homolog (ALG12), encoded on chromosome 22.

The remaining three sequences, CYB_HUMAN, COX1_HUMAN and APOA_HUMAN, proved to be false positives. CYB_HUMAN and COX1_HUMAN are encoded by the human mitochondrial genome, however, in the BLAT search they did not give a perfect match with this genome because the BLAT search uses the nuclear genetic code (not the mitochondrial genetic code). On the other hand, these protein sequences gave shorter, perfect matches with nuclear chromosomes containing sequences related to these genes. To eliminate this source of false positives, in all subsequent studies of Conflict 5 we used a modified BLAT protocol that removed entries encoded by mitochondrial genes (see Materials and Methods).

Analysis of APOA_HUMAN illustrates the point that the primary reason why some sequences may be incorrectly identified by MisPred as chimeras is that there are closely related genes (gene segments) on different chromosomes and these may give short matches with the query sequence. The gene for APOA_HUMAN is located on chromosome 6 but a perfect match with its first 16 residues is also given by chromosome 2. Despite this limitation of our approach (that short perfect matches may be given by different chromosomes), the percentage of false positives is very low (0.02%).

MisPred identified 5 rat proteins as suspected to be chimeras. Two of these sequences, TRPC3_RAT and SYJ2B_RAT, were true chimeras, derived from unrelated genes located on different chromosomes.

BLAT searches identified 18 zebrafish proteins as suspected to be chimeras. One sequence proved to be a false positive since it gave a short perfect match with another chromosome. The remaining zebrafish sequences (ACC4A_DANRE, AN13C_DANRE, BCL7B_DANRE, CC104_DANRE, CHMA1A_DANRE, EXT1B_DANRE, FGFR2_DANRE, JAG1A_DANRE, GP175_DANRE, KC15B_DANRE, LDHA_DANRE, PA24A_DANRE, PALD_DANRE, TC1D1_DANRE, UBE2F_DANRE, UBTD1_DANRE, WN10A_DANRE) appeared to be truly “chimeric” inasmuch as they are derived from two or more genomic regions assigned to different chromosomes. It should be noted, however, that none of these proteins are chimeras of unrelated genes, they all encode full-length sequences typical of the given protein family.

The fact that the proportion of suspicious proteins (i.e. chimeras derived from genomic regions assigned to different chromosomes) is significantly higher in zebrafish (1.4%) than in the other vertebrates (0.00-0.08%) might be explained by assuming that, for some reason, formation of chimeric transcripts of homologous genes is more likely to occur in zebrafish than in other

vertebrates. A possible support for this explanation might be found in the fact that the common ancestor of ray-finned fishes experienced a whole genome duplication, therefore fish genomes usually have duplicates of single orthologs found in other vertebrates [3-5]. In principle, the higher number of close homologs may increase the chance of interchromosomal homologous recombination (or transchromosomal transcription) and the subsequent formation of chimeric transcripts and chimeric proteins.

An implicit prediction of this explanation, however, is that transcripts of trans-chromosomal chimeras are likely to be rare, whereas normal transcripts of the parent genes are expected to be more abundant. Our observation that ESTs corresponding to the “chimeric” transcripts were found, but the predicted “non-chimeric” transcripts were missing from EST databases (data not shown) suggests that this explanation of the high rate of true positives in zebrafish may not be valid.

An alternative explanation is that the chromosomal assignment of contigs encoding these zebrafish genes may not always be correct. It is noteworthy in this respect that the chromosomal matches of these chimeric proteins are sometimes (e.g. AN13C_DANRE, CHMA1A_DANRE, LDHA_DANRE) found at the termini of short zebrafish contigs, the termini of the different contigs coincide with the boundaries of the apparent chimera formation.

MisPred identified no mouse, chicken, worm or fruitfly Swiss-Prot proteins as suspected to be chimeras.

Comments on mispredicted EnsEMBL entries identified by MisPred routines as suspicious

Conflict 2

Mispredicted entries include the rat proteins ENSRNOP00000019138 (a Notch protein homolog) and ENSRNOP00000036086 (related to Angiopoietin-1 receptor precursor, TIE-2), the opossum EPH receptor B1 homologs ENSMODP00000021175, ENSMODP00000031819, ENSMODP00000032387, ENSMODP00000032388, ENSMODP00000032389, ENSMODP00000033795, ENSMODP00000033797, ENSMODP00000033802, ENSMODP00000033804, ENSMODP00000035096, ENSMODP00000035099, ENSMODP00000036530, ENSMODP00000036533, ENSMODP00000036535, ENSMODP00000038060 and ENSMODP00000038066, and the chicken Notch-1 homolog ENSGALP00000003735. All of these proteins lack transmembrane helices, whereas their equivalents from the same or related species do possess transmembrane segments.

Similarly, the frog proteins ENSXETP00000020884, encoding an erbB-3 receptor tyrosine-protein kinase, ENSXETP00000037512, encoding a homolog of Notch 2 preprotein, ENSXETP00000038207, ENSXETP00000041110 and ENSXETP00000041112 encoding homologs of Eph receptor EphA4b, lack the transmembrane helix characteristic of orthologs of these transmembrane proteins. ENSXETP00000040601, which corresponds to the frog ortholog of Ephrin receptor A7, also lacks a typical transmembrane helix between its extracellular and intracellular domains. The missing transmembrane sequence could be corrected using frog EST sequences such as EL820950.

The *Danio rerio* protein ENSDARP00000000102 contains extracellular Fz and Kringle domains and a cytoplasmic Pkinase domain but lacks a transmembrane segment. This mispredicted protein is encoded by the same gene as unplugged isoform FL of *Danio rerio*

(AAT07679) that encodes a functional receptor tyrosine kinase (with intact transmembrane helix) related to skeletal muscle receptor tyrosine kinases.

The seven true positives of *Fugu rubripes* include NEWSINFRUP00000149801, a protein related to Receptor-type tyrosine-protein phosphatase kappa precursor, NEWSINFRUP00000165700 and NEWSINFRUP00000175769, proteins related to EPH receptor tyrosine kinases but lack transmembrane helices separating their extracellular and intracellular domains.

Analysis of the remaining four true positive pufferfish sequences has revealed that they lack transmembrane helices for a different reason: they probably arose through *in silico* fusion of distinct, tandem genes encoding extracellular and intracellular proteins. BLAST searches of the human genome with the pufferfish fusion proteins revealed that homologs of its constituent genes are usually not closely linked (either separated by numerous genes or located on different chromosomes) in the human genome.

For example, NEWSINFRUP00000146634 contains an extracellular IGFBP domain and an intracellular signaling PIP5K domain. BLAST searches revealed that the IGFBP domain is derived from an IGF binding protein 6 related gene, whereas the PIP5K domain is encoded by a phosphatidylinositol-4-phosphate-5-kinase related gene. BLAST searches of the human genome with this fusion protein revealed that orthologs of its constituent genes are separated by more than 4 Mb and numerous genes on chromosome 12 of the human genome.

NEWSINFRUP00000155548 contains extracellular CUB and Trypsin domains and an intracellular RhoGAP domain. BLAST searches revealed that the CUB and Trypsin domains are encoded by a trypsin-related gene, whereas the RhoGAP domain is encoded by a FAM13A1_v2 protein related gene. BLAST searches of the human genome with this fusion protein revealed that

homologs of its constituent genes are separated by more than 6 Mb and numerous genes on chromosome 4 of the human genome.

NEWSINFRUP00000161639 contains an extracellular VWD domain and an intracellular signaling Y_phosphatase domain. BLAST searches revealed that the VWD domain is encoded by an otogelin related gene, whereas the Y_phosphatase domain is encoded by protein tyrosine phosphatase receptor type Q related gene. BLAST searches of the human genome with this fusion protein revealed that orthologs of the constituent genes are closely linked (separated by less than 1 Mb) on chromosome 12 in the human genome.

NEWSINFRUP00000161730 contains extracellular PSI domain and intracellular LIM and SH3_1 domains. BLAST searches revealed that the PSI domain is encoded by Plexin domain-containing protein 1 precursor related gene, whereas the LIM and SH3_1 domains are encoded by LIM and SH3 protein 1 related gene. BLAST searches of the human genome with this fusion protein revealed that orthologs of the constituent genes (*LASPI* and *PLXDI*) are also linked (on chromosome 17) in the human genome.

Conflict 3

The MisPred analyses of EnSEMBL sequences for Conflict 3 have identified 1 human, 3 mouse, 3 rat, 0 opossum, 1 chicken, 0 frog, 2 pufferfish, 0 zebrafish, 0 sea squirt, 0 worm and 0 fruitfly EnSEMBL sequences as suspicious, i.e. they contain both extracellular and nuclear domains.

Human ENSP00000374642, mouse ENSMUSP00000023060, mouse ENSMUSP00000086712, mouse ENSMUSP00000086714, rat ENSRNOP00000004703, rat ENSRNOP00000022586, rat ENSRNOP00000052999, and chicken ENSGALP00000019866 were all found to contain extracellular Pentaxin and nuclear Chromo domains. BLAST searches

revealed that their Pentaxin domains are encoded by a neuronal pentraxin receptor related gene, whereas the Chromo domain is encoded by a Chromobox protein homolog 6 related gene.

The *Fugu rubripes* protein NEWSINFRUP00000161053 contains nuclear SNF2_N and Helicase_C domains (intracellular Ras domains) and extracellular Cadherin domains. BLAST searches revealed that the SNF2_N and Helicase_C domains are encoded by a homolog of DNA repair and recombination protein RAD54B (the Ras domain is encoded by a gene related to the GTP binding protein gene overexpressed in skeletal muscle) and the Cadherin domains are encoded by a Cadherin 17 related gene. BLAST searches of the human genome with this pufferfish fusion protein revealed that orthologs of the three constituent genes (*RAD54b*, *GEM* and *CDH17*) are also linked (on chromosome 8) in the human genome.

The *Fugu rubripes* protein NEWSINFRUP00000168199 contains nuclear zf-C3HC4 and zf-B_box domains (an intracellular SPRY domain) and an extracellular C1q domain. BLAST searches revealed that the zf-C3HC4, zf-B_box (and SPRY) domains are encoded by a RING finger protein-related gene, whereas the C1q domain is encoded by a cerebellin 1 precursor related gene. BLAST searches of the human genome with this fusion protein revealed that homologs of its constituent genes are not linked in the human genome.

Comments on erroneous GNOMON-predicted proteins identified by MisPred routines as suspicious

Conflict 2

Proteins XP_001372663, XP_001372858 and XP_001381008 contain both extracellular (SEA) and intracellular (C1_1, Pkinase, Pkinase_C) domains but lack transmembrane helices. BLAST searches have confirmed that these predicted protein sequences arose through the fusion of genes encoding a protein related to mucin 16 and a gene encoding a protein kinase C zeta type. BLAST searches of the human genome revealed that the constituent genes are located on different chromosomes (chromosome 19 and chromosome 1).

Protein XP_426670 contains an extracellular Insulin domain and intracellular WD40 domains, but no transmembrane domain. BLAST searches revealed that this predicted protein sequence arose through the fusion of tandem genes encoding a protein related to WD repeat protein and a relaxin 3 precursor related gene. BLAST searches of the human genome revealed that the constituent genes are located on different chromosomes (chromosome 1 and chromosome 19) in the human genome.

Protein XP_424494 contains an extracellular Somatomedin_B domain and intracellular RasGEF and DEP domains but no transmembrane segment. BLAST searches revealed that this predicted protein sequence arose through the fusion of tandem genes encoding a protein related to placental protein 11 and a Rap guanine nucleotide exchange factor (GEF) 3 related gene. BLAST searches of the human genome with this chicken fusion protein revealed that the constituent genes are also closely linked (on chromosome 12) in the human genome. The fact that no cDNAs or ESTs were found that span these constituent genes, suggests that they are distinct genes.

XP_001337907 was found to contain extracellular Lectin_C and intracellular kinesin domains. BLAST searches revealed that the protein arose through fusion of a kinesin-related gene and a novel C-type lectin domain containing protein.

XP_001340361 contains extracellular SRCR domains and an intracellular CH domain. This protein apparently arose through fusion of a Scavenger receptor cysteine-rich type 1 protein-related gene and a gene encoding a filamin-related protein. BLAST searches revealed that the constituent genes are located on different chromosomes (chromosome 12 and chromosome X) in the human genome.

XP_696912 contains an extracellular Laminin_G_2 domain and an intracellular Exo_endo_phos domain. This protein apparently arose through fusion of a collagen XXVII proalpha 1 chain precursor related gene and an inositol polyphosphate-5-phosphatase E related gene. BLAST searches of the human genome with this zebrafish fusion protein revealed that the constituent genes are on chromosome 9 in the human genome separated by more than 22 Mb (and numerous genes).

XP_001343892 contains extracellular EGF domains and an intracellular VPS9 domain. This protein apparently arose through fusion of a NOTCH related gene and a Rab5-activating protein 6 related gene. BLAST searches of the human genome with this zebrafish fusion protein revealed that the constituent genes are on chromosome 9 in the human genome, separated by 10 Mb (and numerous genes).

XP_690248 contains two extracellular CUB domains and an intracellular CH domain. BLAST searches revealed that the CUB domains are encoded by a procollagen C-endopeptidase enhancer 2 related gene, whereas the intracellular CH domain is encoded by the plastin 1 gene. BLAST searches of the human genome with this zebrafish fusion protein revealed that the constituent genes are linked on chromosome 3 in the human genome separated by a gene for

transient receptor potential cation channel, subfamily C, member 1. EST searches provided no evidence for cotranscription of these genes.

XP_001334415 contains a C-terminal extracellular OLF domain and an N-terminal intracellular signaling WH1 domain. BLAST searches revealed that the OLF domain is encoded by latrophilin-2 related gene, the WH1 domain is encoded by Wiskott-Aldrich syndrome-related gene, whereas the middle of the protein is related to retroviral polyproteins. BLAST searches of the human genome with this zebrafish fusion protein revealed that the constituent genes are on two chromosomes in the human genome (Wiskott-Aldrich syndrome-related gene is on chromosome 7, latrophilin-2 related gene is on chromosome 19).

Conflict 3

XP_001375906 was found to contain both extracellular (MHC_I, C1_set) and nuclear (Pou, Homeobox) domains. BLAST searches have confirmed that this predicted protein sequence arose through the fusion of tandem genes encoding the opossum ortholog of the POU domain, class 5, transcription factor, PO5F1_PIG, and a gene encoding a protein related to MHC class I antigens. BLAST searches of the human genome with this opossum fusion protein revealed that the constituent genes are on different chromosomes (chromosome 12 and chromosome 6) in the human genome.

XP_001380072 was found to contain three extracellular WAP domains and a nuclear histone domain. BLAST searches revealed that this predicted protein arose through fusion of a gene encoding a whey acidic protein-like protein and a gene for Histone H2A.1. BLAST searches of the human genome with this opossum fusion protein revealed that homologs of the constituent genes are located on different chromosomes in the human genome.

XP_425476 contains an extracellular Pentaxin and a nuclear Chromo domain. BLAST searches revealed that the Pentaxin domain is encoded by a neuronal pentraxin receptor related gene, whereas the Chromo domain is encoded by a Chromobox protein homolog 6 related protein. As discussed above (cf. EnSEMBL sequences) homologs of the constituent genes are closely linked in human and other mammalian species.

XP_686306 contains an extracellular Lectin_C and a nuclear zf-B_box domain. BLAST searches revealed that the Lectin_C domain is encoded by gene related to serum lectin isoform 5 precursor of *Salmo salar*, whereas the zf-B_box domain is encoded by a Tripartite motif-containing protein 2 related protein. BLAST search provided no evidence for the linkage of related genes in the human genome.

XP_001337208 contains extracellular SRCR domain and nuclear SCAN and zf-CCHC domains. BLAST searches revealed that the SCAN and zf-CCHC domains are encoded by a gene related to LReO_3 of *Oryzias latipes*, whereas the SRCR domain is most closely related to those of neurotrypsin. BLAST searches revealed that homologs of the constituent genes are not linked in the human genome.

XP_001338835 contains an extracellular A4_EXTRA domain and nuclear zf-C2H2 domains. BLAST searches revealed that the zf-C2H2 domains are encoded by the gene encoding the zebrafish protein CAM13007, similar to vertebrate PR domain containing 10 (PRDM10). The A4 EXTRA domain is encoded by the gene encoding the zebrafish protein CAM13008, a novel protein similar to vertebrate amyloid beta (A4) precursor-like protein 2 (APLP2). BLAST searches of the human genome with this fusion protein revealed that the constituent genes are closely linked (separated by less than 1 Mb) on chromosome 11 in the human genome. Note, however, that no ESTs were found that would support co-transcription of these genes in vertebrate genomes.

XP_001332478 contains extracellular Ldl_recept_a and Ldl_recept_b domains and a nuclear zf-C4 domain. BLAST searches revealed that the zf-C4 domain is encoded by a gene related to hepatocyte nuclear factor 4 alpha isoform, whereas the Ldl_recept_a and Ldl_recept_b domains are most closely related to low-density lipoprotein receptor-related protein 1 precursor. BLAST searches revealed that orthologs of the constituent genes are located on different chromosomes (chromosome 20 and chromosome 12) in the human genome.

Comments on EnsEMBL and NCBI entries predicted for the same set of genes

Conflict 1

MisPred analyses of EnsEMBL and NCBI entries predicted for the same set of genes have revealed that the proportion of suspicious sequences detectable by the routine for Conflict 1 are quite similar for EnsEMBL- and GNOMON-predictions in the case of human (15.65% vs. 23.02%), opossum (20.15% vs. 17.45%) and chicken (32.75% vs. 28.39%) genes. In the case of zebrafish genes, however, the proportion of suspicious proteins was found to be lower in the case of GNOMON-predicted proteins (21.29%) than EnsEMBL entries (36.86%).

Conflict 2

MisPred analyses have returned no human EnsEMBL or GNOMON-predicted sequences for Conflict 2 as being erroneous. In the case of chicken, MisPred analyses have returned 1 GNOMON protein but no EnsEMBL proteins as suspicious. The GNOMON-predicted protein was a false positive due to lack of detection of its transmembrane helix by TMHMM.

In the case of opossum no EnsEMBL sequences but four GNOMON-predicted sequences were identified by MisPred as suspicious. Two of the latter were false positives due to lack of detection of their transmembrane helices by TMHMM, but two were truly erroneous: they arose through *in silico* fusion of genes encoding extracellular and intracellular proteins.

In the case of *Danio rerio*, MisPred analysis identified five EnsEMBL sequences and six GNOMON-predicted sequences as being suspicious. Four of the EnsEMBL entries and four of the GNOMON-predicted entries were false positives due to failure of TMHMM to detect their transmembrane helices or false Pfam-A domain hits.

One truly erroneous EnsEMBL sequence, ENSDARP00000000102, similar to Tyrosine-protein kinase transmembrane receptor ROR1 precursor, contains extracellular Fz and Kringle domains and an intracellular Pkinase domain but lacks the transmembrane segment separating its constituent extracellular and intracellular domains. It should be noted that the corresponding GNOMON-predicted sequence, XP_001332675, does contain the expected transmembrane helix.

The remaining two truly erroneous GNOMON-predicted proteins of *Danio rerio* apparently arose through the *in silico* fusion of genes encoding extracellular and intracellular proteins. XP_690248 arose through fusion of a procollagen C-endopeptidase enhancer 2 related gene and a plastin 1 related gene. Similarly, XP_001334415 probably arose through fusion of a latrophilin-2 related gene and a Wiskott-Aldrich syndrome related gene.

In summary, the results of the MisPred analyses of human, opossum, chicken and zebrafish sequences suggest that the GNOMON and EnsEMBL pipelines are characterized by similarly low proportion of error detectable by the routine for Conflict 2. It should be pointed out, however, that GNOMON appears to be slightly more prone to erroneous fusion of genes.

Conflict 3

MisPred analysis has found no human or chicken EnsEMBL or GNOMON-predicted proteins containing both extracellular and nuclear Pfam-A domains.

In the case of opossum EnsEMBL proteins no sequence was found to contain both extracellular and nuclear Pfam-A domains, whereas in the case of opossum GNOMON-predicted proteins XP_001375906 contains an extracellular MHC-I domain and intracellular signaling Pou and Homeobox domains. This protein probably arose through the *in silico* fusion of Hereditary hemochromatosis protein homolog precursor related gene and a POU domain class 5 transcription factor 1-like protein related gene.

MisPred analysis of zebrafish entries has found no EnSEMBL proteins containing both extracellular and nuclear Pfam-A domains, whereas two GNOMON-predicted proteins were identified as suspicious by MisPred. One of these sequences was a false positive, due to a false Pfam-A domain hit. Protein XP_001332478 was found to contain extracellular Ldl_recept_a and Ldl_recept_b domains and an intracellular signaling zf-C4 domain. This sequence probably arose through the *in silico* fusion of a Hnf4a protein related gene and a gene closely related to low-density lipoprotein receptor-related protein 1.

In summary, MisPred analysis of human, opossum, chicken and zebrafish sequences predicted by GNOMON and EnSEMBL pipelines has revealed that the rates of error detectable by Conflict 3 are characterized by similarly very low values for all of these species. It should be pointed out, however, that GNOMON appears to be slightly more prone to the erroneous fusion of tandem genes.

Conflict 4

MisPred analyses of EnSEMBL and GNOMON entries predicted for the same set of genes have revealed that the proportion of suspicious sequences predicted by the two pipelines are not drastically different for human (6.65% vs. 5.18%) and chicken (9.70% vs. 13.87%) genes, but there are somewhat greater differences regarding opossum (3.19% vs. 1.14%) and zebrafish (7.68% vs. 14.57%) genes.

Conflict 5

MisPred identified none of the human EnSEMBL- and only one of the GNOMON-predicted sequences as being encoded by two genomic segments located on different chromosomes. As discussed above XP_001128605 (now replaced by the nonchimeric NP_001035225) is a chimera of genomic regions located on chromosomes 2 and 7.

In the case of zebrafish, Mispred identified none of the EnsEMBL- and five of the GNOMON-predicted sequences as suspicious. Analyses of the latter sequences have shown that none of them are truly chimeric: they had >90 % full-length match with chromosome regions designated as NA_random or UN_random.

Part 2. Detailed description of the detection of suspicious proteins by Conflict 4

Identification of Pfam-A families suitable for the study of the domain integrity

In order to identify the Pfam-A domain families that have a well-defined, conserved sequence length range (and are thus suitable for the study of domain integrity), we selected only those families whose members do not deviate from the average size by more than 2 SD values in the high quality Swiss-Prot database (version 48.9).

First, we calculated the average length and standard deviation for all Pfam-A domain families.

For this calculation we used the seed alignments provided by Pfam (Pfam Release 19.0) since the sequences included in these alignments are usually full-length representatives of the Pfam-A domain families. Pfam Release 19.0 was downloaded from

<http://www.sanger.ac.uk/Software/Pfam/index.shtml>.

In the next step we analyzed Pfam-A domains in Swiss-Prot sequences and selected those presumably incomplete Swiss-Prot domains whose beginning or end was missing their first or last >5 positions in the full alignments of the domain families, and whose lengths differed by at least 2 standard deviations from the Pfam-A seed averages. To check if these domains are genuinely incomplete or their domain boundaries were incorrectly defined by Pfam we ran a blastp search with the Swiss-Prot proteins containing the incomplete domains as queries against the full set of the Pfam-A domains to identify the missing parts of the domains. If the length of the corrected domain was inside the conserved length range (seed average ± 2 SD) then the domain was considered to be complete.

Pfam-A domain families were deemed to be suitable for the study of domain integrity if they have either never been incomplete in Swiss-Prot proteins or all members of the family could be completed. Based on these criteria, about 90% of the Pfam-A domain families in the Swiss-

Prot proteins proved to have a well-defined, conserved length range and so can be used to identify mispredicted or abnormal proteins containing domains of deviant length. Furthermore, we also found that many Pfam-A domains that appear fragments at first sight in the seed alignments can be easily amended into full-length domains by comparing the vicinity of these domain fragments to homologous full-length Pfam-A domains. Of the 2752 Pfam-A domain families present in human Swiss-Prot proteins 2529 families, of the 2779 Pfam-A domain families present in vertebrate Swiss-Prot proteins 2515; of the 3559 Pfam-A domain families present in metazoa+fungi Swiss-Prot proteins 3010 families appeared to be suitable for the study of domain integrity.

Interestingly, a survey of the Pfam-A families that most frequently violate domain integrity revealed that they are usually found in numerous tandem copies in proteins (e.g. collagen triple helix repeat, Ankyrin repeat, Spectrin repeat, Plectin repeat, Tetratricopeptide repeat, cadherin domains, fibronectin type III domain, etc.) (see Supplementary table 1). The presence of the collagen triple helix repeat in this table simply reflects the fact that the dogma of a narrow domain-size range is not valid for fibrous proteins such as collagens.

In order to avoid artificial split-up of domain matches by blastp, we did not mask low complexity regions. However, in some instances this resulted in partial matches of two domains belonging to different Pfam-A families, therefore we removed these matching Pfam-A families from the set of the reliable Pfam-A families.

When we compared the domains of the ‘reliable’ Pfam-A families to one another we identified several pairwise matches with high significance and pairwise similarity (>60% identity) where the two domains belong to different Pfam-A families (see Supplementary table 2). Many of these Pfam-A domains match one another with high similarity because one domain is embedded in another. For example, the most frequently co-occurring pairs are the Fibronectin

type II domains, fn2 (PF00040) and peptidase_M10 (PF00413) domains, where the PF00040 was found to be embedded in a PF00413 domain 766 times (in gelatinases). As overlapping domains would cause ambiguity in the filtering of deviant domains, we eliminated the longer Pfam-A families that encompass the shorter ones. Another reason why members of different Pfam-A families may match one another with high similarity is because they belong to different but closely related Pfam families. For example, both bZIP_2 and bZIP_1 belong to the bZIP-like leucine zipper clan, both Arm and HEAT belong to the Armadillo repeat superfamily clan, both LRR_1 and LRR_2 belong to the Leucine Rich Repeat clan, both APH and Choline_kinase belong to the Protein kinase superfamily clan (see Supplementary table 2). To avoid ambiguity in the filtering of deviant domains, we eliminated both Pfam-A families from the list of domains suitable for the filtering task.

Using this approach we created databases of human, vertebrate and metazoa+fungi Swiss-Prot domain sequences belonging to Pfam-A families suitable for the study of domain integrity. The list of Pfam-A domain families suitable for the study of domain integrity is deposited in Additional file 6.

Identification of cases of violation of the domain integrity

In order to set a proper similarity threshold for the domain searches above which most partial matches would be eliminated we ran an all-against-all blastp search for the domains of all reliable Pfam-A families (i.e. whose members were never incomplete or where all could be repaired). The results for the human Sushi domains are shown in Supplementary figure 1 where the relative lengths (shorter/longer) of any two matching domains are plotted as a function of their percentage identity. As a contrast we also show the results of an all-against-all blastp search for all human Cadherin domains (see Supplementary figure 2). Unlike the conserved-length Sushi domains, the relative lengths of the matches among the Cadherin domains vary greatly even

above 60% identity, the threshold suggested above. It must be noted that as Cadherin proved to be a Pfam-A family of irregular type where not all of the incomplete members could be repaired; this domain type was already disqualified as a filter to search for incorrectly defined domains (see above).

After running an all-against-all blastp search for all domains we were able to establish the reliability of the approach as a function of the required sequence similarity between any two matching domains of the same type. We found that using a cut-off value of 60% sequence identity of two domain sequences eliminated at least 99% of partial matches, thus reducing the number of false positives to below 1%, i.e. this is an appropriate cut-off value for pairwise domain comparisons to filter for proteins containing fragment domains.

We ran a blastp search with the current set of proteins as queries against the appropriate reliable Swiss-Prot domain sequences, and selected those partial domain matches which share over 60% identity with the query sequence, with an E-value $< 1e^{-5}$, and differ by at least 40% in length. The protein sequences containing these domains with deviant lengths were identified as suspicious.

Size-distribution of gaps in pairwise matches of Pfam-A domains

We calculated the longest continuous gap for each pairwise domain match with $> 60\%$ sequence identity. The longest gap (indel) for any pairwise match with $>60\%$ sequence identity between any two domains belonging to a reliable Pfam-A family was 51, whereas between any two domains belonging to an unreliable Pfam-A family it was 52. The frequency distribution of the gaps by gap-size for the two sets of domains is shown in Supplementary figure 3.

It is clear from this figure that the data closely follow straight lines in double-logarithmic plots for both the reliable and the unreliable domain families. What is more, there is no significant difference between the two slopes. In other words, gap size distribution follows a

power law for both groups, i.e. the shortest gaps are the most abundant, whereas longer gaps are increasingly less frequent. This type of distribution is consistent with (and supports) the dogma on which MisPred routine 4 is based: insertion/deletion of longer segments are increasingly less likely to be compatible with the viability of a protein fold.

References

1. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion.** *Protein Eng Des Sel.* 2004, **17**:349-56.
2. Qi H, Gervais ML, Li W, DeCaprio JA, Challis JR, Ohh M: **Molecular cloning and characterization of the von Hippel-Lindau-like protein.** *Mol Cancer Res.* 2004, **2**:43-52.
3. Robinson-Rechavi M, Marchand O, Escriva H, Bardet PL, Zelus D, Hughes S, Laudet V: **Euteleost fish genomes are characterized by expansion of gene families.** *Genome Res.* 2001, **11**:781-8.
4. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y: **Genome duplication, a trait shared by 22000 species of ray-finned fish.** *Genome Res.* 2003, **13**:382-90.
5. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-I T, Takeda H, Morishita S, Kohara Y: **The medaka draft genome and insights into vertebrate genome evolution.** *Nature* 2007, **447**:714-9.

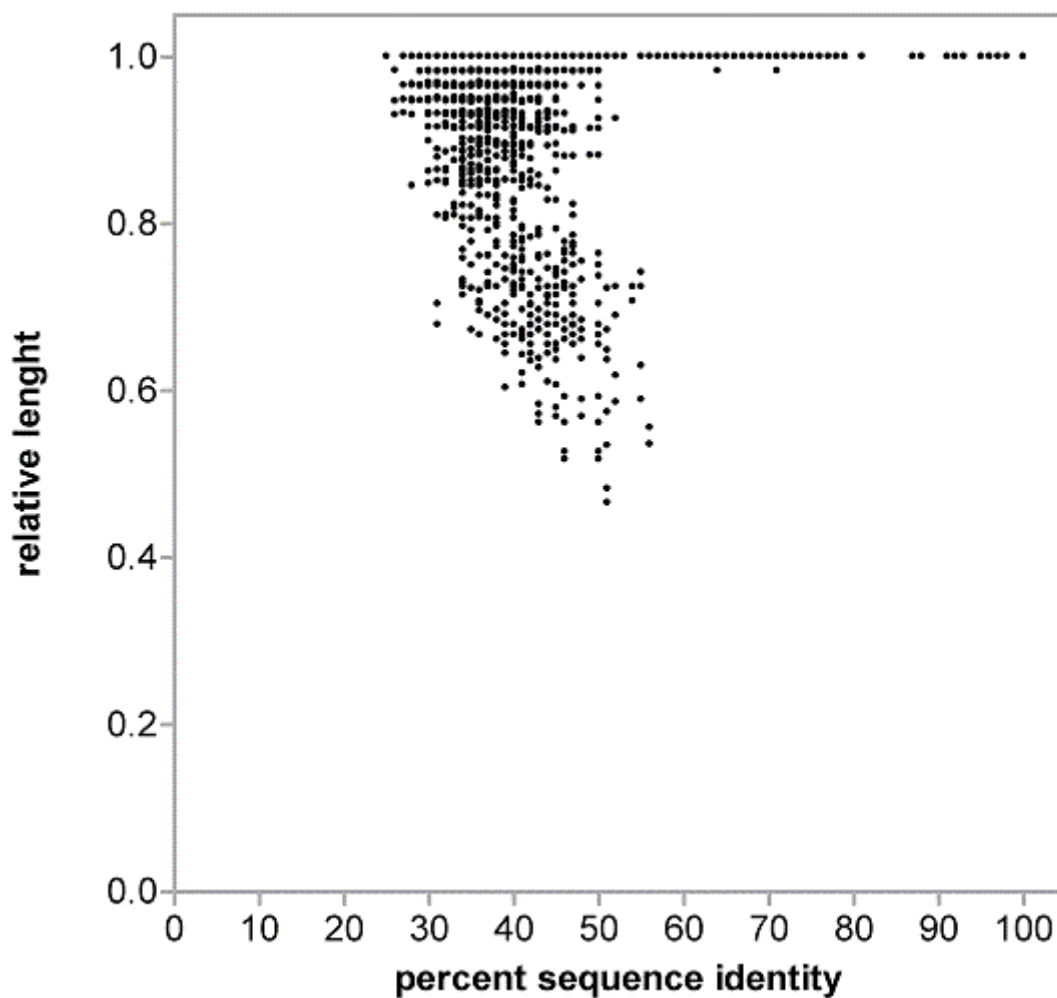
Supplementary table 1. Pfam-A families that most frequently violate domain integrity in human Swiss-Prot proteins.

Pfam ID	Number of copies in human Swiss-Prot entries	Number of deviant family members	Percent of deviant family members	Number of deviant family members that could be corrected	Percent of deviant family members that could be corrected	Number of deviant family members that could not be corrected	Percent of deviant family members that could not be corrected	Average length	Standard deviation	Description
PF01391	487	192	39	185	96	7	4	60	0	Collagen triple helix repeat (20 copies)
PF00023	717	107	15	30	28	77	72	33,48	2,35	Ankyrin repeat
PF00028	630	97	15	90	93	7	7	94,53	4,81	Cadherin domain
PF00400	688	54	8	30	56	24	44	39,23	2,35	WD domain, G-beta repeat
PF00076	224	36	16	24	67	12	33	70,6	2,53	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00435	315	32	10	19	59	13	41	105,85	2,71	Spectrin repeat
PF00041	402	27	7	16	59	11	41	84,44	5,47	Fibronectin type III domain
PF01344	171	24	14	12	50	12	50	47,22	3,06	Kelch motif
PF00053	167	24	14	12	50	12	50	50,67	5,07	Laminin EGF-like (Domains III and V)
PF00069	302	24	8	20	83	4	17	270,75	17,3	Protein kinase domain
PF00681	76	19	25	9	47	10	53	44,95	0,3	Plectin repeat
PF00595	202	18	9	10	56	8	44	84,1	5,01	PDZ domain (Also known as DHR or GLGF)
PF00018	151	16	11	13	81	3	9	56,11	1,09	SH3 domain
PF00071	107	14	13	11	79	3	21	165,21	6,36	Ras family
PF00036	303	13	4	4	31	9	69	28,89	0,56	EF hand
PF00063	38	12	32	2	17	10	83	678,32	17,52	Myosin head (motor domain)
PF00989	19	12	63	5	42	7	58	111,55	5,12	PAS fold
PF00067	57	12	21	8	67	4	33	459,12	11,37	Cytochrome P450
PF00324	21	12	57	12	100	0	0	467,64	23,01	Amino acid permease
PF00515	247	11	4	3	27	8	73	34	0,07	Tetratricopeptide repeat

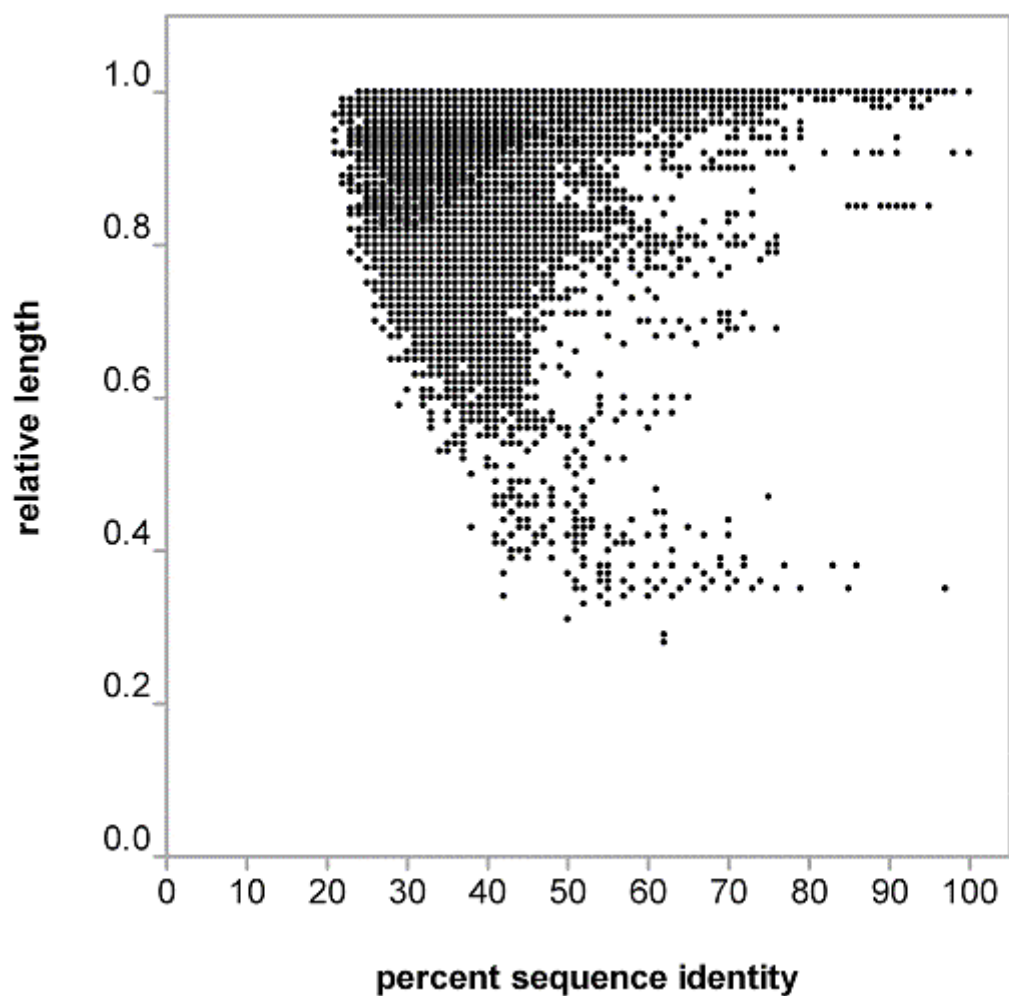
Supplementary table 2. Nested and overlapping Pfam-A domain families.

Pfam1: Pfam ID of longer domain	Pfam name	Average size	Standard deviation	Pfam2: Pfam ID of shorter domain	Pfam name	Average size	Standard deviation	Number of times Pfam1 and Pfam2 overlap
PF00413	Peptidase_M10	164,59	31,97	PF00040	fn2	41,34	1,28	766
PF00406	ADK	177,13	12,95	PF05191	ADK_lid	36,06	0,32	97
PF00389	2-Hacid_dh	317,81	8,52	PF02826	2-Hacid_dh_C	178,23	5,4	89
PF02463	SMC_N	1033,91	307,22	PF06470	SMC_hinge	116,73	7,95	89
PF07716	bZIP_2	54,53	1,16	PF00170	bZIP_1	58,94	2,31	57
PF05221	AdoHcyase	457,38	21,27	PF00670	AdoHcyase_NAD	162,25	1,02	40
PF00478	IMPDH	446,25	54,33	PF00571	CBS	123,73	15,22	25
PF04857	CAF1	305,75	122,27	PF01424	R3H	56,08	1,64	25
PF00443	UCH	417,91	131,18	PF00627	UBA	40,49	0,82	18
PF05192	MutS_III	305,26	9,61	PF05190	MutS_IV	93,63	2,43	18
PF03917	GSH_synth_ATP	479,4	10,82	PF03199	GSH_synthase	104,55	2,16	17
PF02166	Androgen_recep	436,75	9,22	PF02155	GCR	374,5	2,12	12
PF01546	Peptidase_M20	322,42	24,47	PF07687	M20_dimer	115,06	21,64	9
PF02463	SMC_N	1033,91	307,22	PF04423	Rad50_zn_hook	55,21	1,18	9
PF01137	RTC	329,09	9,34	PF05189	RTC_insert	101,9	6,48	8
PF01193	RNA_pol_L	195,56	68,81	PF01000	RNA_pol_A_bac	120,27	8,7	8
PF00514	Arm	41,26	1,21	PF02985	HEAT	38,72	1,01	5
PF07723	LRR_2	25,92	1,03	PF00560	LRR_1	22,97	2,17	5
PF04998	RNA_pol_Rpb1_5	580,84	52,22	PF04990	RNA_pol_Rpb1_7	134,29	1,99	4
PF04998	RNA_pol_Rpb1_5	580,84	52,22	PF04992	RNA_pol_Rpb1_6	189,43	8,17	4
PF07690	MFS_1	368,24	29,28	PF05978	DUF895	156,27	1,42	4
PF03881	Fructosamin_kin	287,5	1	PF01636	APH	236,9	16,93	3
PF04931	DNA_pol_V	768,8	28,02	PF03066	Nucleoplasmin	175,68	31,07	3
PF01500	Keratin_B2	153,44	20,49	PF04360	Serglycin	150	0	2
PF01636	APH	236,9	16,93	PF01633	Choline_kinase	219,12	23,36	2
PF03344	Daxx	737,5	2,12	PF04849	HAP1_N	306	13,66	2
PF05308	DUF729	240,25	14,8	PF06346	Drf_FH1	161,5	10,28	1

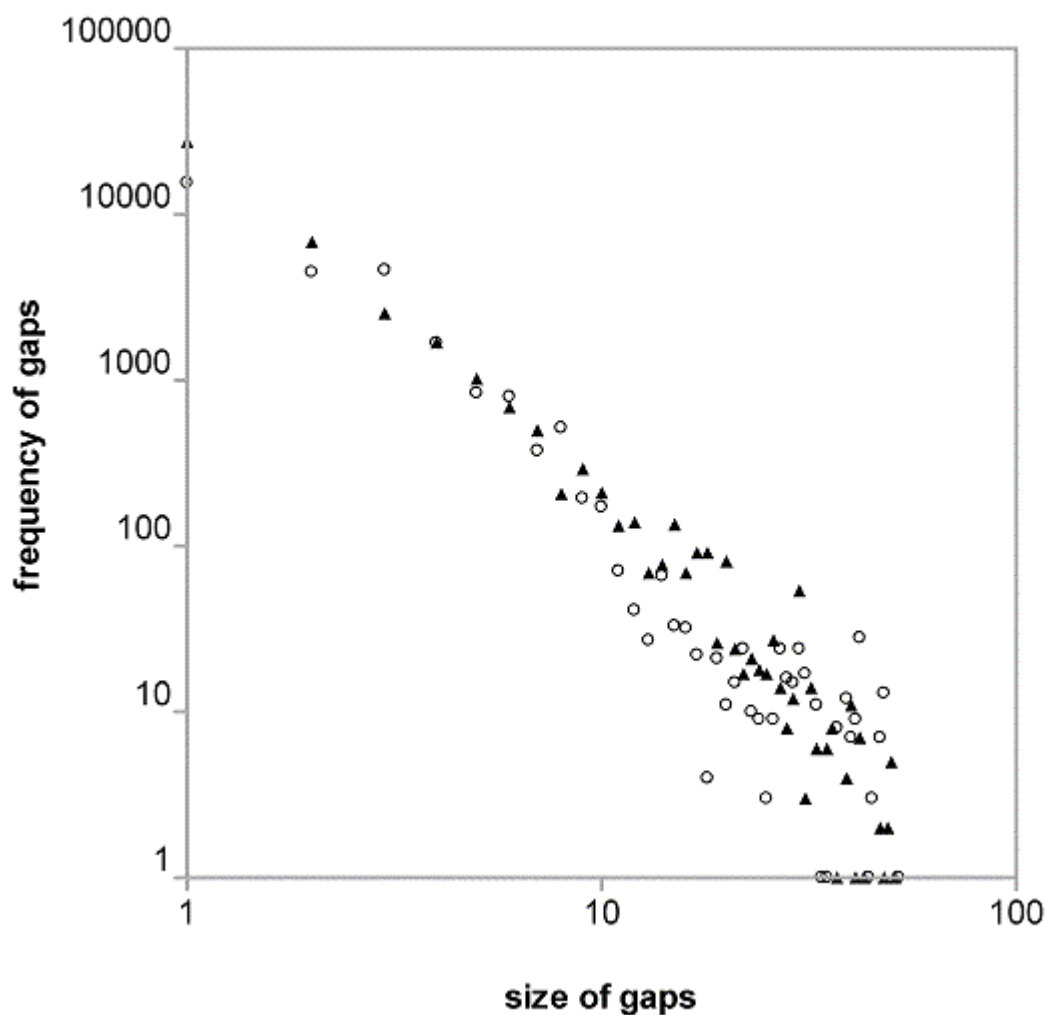
Supplementary figure 1. Relative lengths of 254 matching human Sushi domains in pairwise blastp alignments. Relative length (calculated by dividing the length of the shorter with that of the longer member of the pairs aligned) is plotted as a function of percent sequence identity. Note that the length of matching domains does not deviate by more than 10% if sequence identity is >60%.



Supplementary figure 2. Relative lengths of 630 matching human Cadherin domains in pairwise blastp alignments. Relative length (calculated by dividing the length of the shorter with that of the longer member of the pairs aligned) is plotted as a function of percent sequence identity. Note that the length of matching domains deviates by more than 10% even if sequence identity is >95%. This domain family was not used for the filtering process described in this paper as not all its deviant domain representatives could be repaired.



Supplementary figure 3. Frequency distribution of gaps for regular (▲) and deviant (○) domain families with respect to gap-size in pairwise domain alignments. The longest continuous gap was considered for any pairwise domain alignment with >60% percentage identity. The figure shows the frequency of gaps of a given size (y axis, log10 scale) as a function of the size of the gaps (x axis, log10 scale). Note that the longest recorded indel was 51 and 52 amino acids for regular and deviant domain families, respectively. The linear nature of the log-log representation of the data reveals a power-law distribution.



Additional file 2. List of erroneous Swiss-Prot sequences identified by MisPred.
Comments on these Swiss-Prot entries are found in Additional file 1, Part 1. Entries not annotated as fragments or chimeras in Swiss-Prot are in bold in the table.

<i>Homo sapiens</i>	Conflict 1	Conflict 4	Conflict 5
	ACROL_HUMAN	HB2T_HUMAN	CR030_HUMAN
	MUC5A_HUMAN	POK9_HUMAN	MED12_HUMAN
	POZP3_HUMAN	M4A4E_HUMAN	
	HB2G_HUMAN	VHLL_HUMAN	
	LPLC4_HUMAN		
<i>Mus musculus</i>	Conflict 1	Conflict 4	Conflict 5
	C209C_MOUSE	MYO1A_MOUSE	
	LYG6_MOUSE		
	MCPT3_MOUSE		
	NOE2_MOUSE		
	TMPS7_MOUSE		
<i>Rattus norvegicus</i>	Conflict 1	Conflict 4	Conflict 5
	ANGP2_RAT	ADA18_RAT	TRPC3_RAT
	CADH4_RAT	AL7A1_RAT	SYJ2B_RAT
	CATG_RAT	AMPD2_RAT	
	CO5_RAT	CO8B_RAT	
	CO8B_RAT	DUS7_RAT	
	COBA1_RAT	EPHA5_RAT	
	CSF2_RAT	E2F1_RAT	
	FA9_RAT	FER_RAT	
	GDF11_RAT	KITH_RAT	
	KLK10_RAT	MAAI_RAT	
	KLK3_RAT	PTN23_RAT	
	MIP2A_RAT	RPGF4_RAT	
	MUC2L_RAT	SFRP1_RAT	
	NID1_RAT	UCK2_RAT	
	PDGFB_RAT		
	SFRP1_RAT		
<i>Gallus gallus</i>	Conflict 1	Conflict 4	Conflict 5
	SECR_CHICK	ALDOC_CHICK	
	CO2A1_CHICK	AMPD1_CHICK	
	TGFB1_CHICK	AT2B1_CHICK	
	CATL_CHICK	FZD3_CHICK	
	FINC_CHICK	FZD6_CHICK	
	PYY_CHICK	HXA9_CHICK	
	IGF2_CHICK	LAMBV_CHICK	
	NMU_CHICK	S100G_CHICK	
	AMP1_CHICK		
	CFBL_CHICK		
	RNL2_CHICK		
	IPK1L_CHICK		
	LAMB1_CHICK		
	LAMBV_CHICK		
	FIBB_CHICK		
	SEM4D_CHICK		
	BMP2_CHICK		
	WNT1_CHICK		
	SORL_CHICK		
<i>Danio rerio</i>	Conflict 1	Conflict 4	Conflict 5
	OSTC_BRARE	DHH_BRARE	
<i>Caenorhabditis elegans</i>	Conflict 1	Conflict 4	Conflict 5
	YOW6_CAEEL	PME6_CAEEL	
	YL54_CAEEL	YQS2_CAEEL	
	YL15_CAEEL		
	NAS10_CAEEL		
	LAML1_CAEEL		
	YP95_CAEEL		
	NAS13_CAEEL		
	CUBN_CAEEL		
<i>Drosophila melanogaster</i>	Conflict 1	Conflict 4	Conflict 5

	CHIT3_DROME		
	CHIT1_DROME		

Additional file 3. List of extracellular Pfam-A domain families. The table contains obligatory extracellular Pfam-A domain families which were used to predict subcellular localization of proteins. Our domain co-occurrence analyses of Metazoan UniProtKB entries have identified 166 obligatory extracellular Pfam-A domain families, the majority of which are also identified as such in the SMART database.

Description	SMART name	Pfam name	Pfam ID
Amyloid A4	A4_EXTRA	A4_EXTRA	PF02177
Plant lipid transfer protein / seed storage protein / trypsin-alpha amylase inhibitor domain family	AAI	Tryp_alpha_amyl	PF00234
Serum albumin	ALBUMIN	Serum_albumin	PF00273
Ami_2	Ami_2	Amidase_2	PF01510
Ami_3	Ami_3	Amidase_3	PF01520
Adhesion-associated domain present in MUC4 and other proteins	AMOP	AMOP	PF03782
Anaphylatoxin homologous domain	ANATO	ANATO	PF01821
APPLE domain	APPLE	PAN	PF00024
Bulb-type mannose-specific lectin	B_lectin	B_lectin	PF01453
Bacterial Ig-like domain (group 1)	BID_1	Big_1	PF02369
Bacterial Ig-like domain 2	BID_2	Big_2	PF02368
Bacterial OsmY and nodulation domain	BON	BON	PF04972
Bowman-Birk type proteinase inhibitor	BowB	Bowman-Birk_leg	PF00228
BPI/LBP/CETP N-terminal domain	BPI1	LBP_BPI_CETP	PF01273
Complement component C1q domain	C1Q	C1q	PF00386
Netrin C-terminal Domain	C345C	NTR	PF01759
C-terminal tandem repeated domain in type 4 procollagens	C4	C4	PF01413
Cadherin repeats	CA	Cadherin	PF00028
Calcitonin	CALCITONIN	Calc_CGRP_IAPP	PF00214
Cellulose Binding Domain Type IV	CBD_IV	CBM_6	PF03422
Domain abundant in complement control proteins; SUSHI repeat; short complement-like repeat (SCR)	CCP	Sushi	PF00084
CFEM	CFEM	CFEM	PF05730
A domain in the BMP inhibitor chordin and in microbial proteins	CHRD	CHRD	PF07452
Chitin binding domain	ChtBD1	Chitin_bind_1	PF00187
Chitin-binding domain type 2	ChtBD2	CBM_14	PF01607
Chitin-binding domain type 3	ChtBD3	CBM_5_12	PF02839
Clostridial hydrophobic, with a conserved W residue, domain	ChW	ChW	PF01584
CLUSTERIN alpha chain	CLa	Clusterin	PF01093
CLUSTERIN Beta chain	CLb	Clusterin	PF01093
C-type lectin (CTL) or carbohydrate-recognition domain (CRD)	CLECT	Lectin_C	PF00059
Connexin homologues	CNX	Connexin	PF00029
Fibrillar collagens C-terminal domain	COLFI	COLFI	PF01410
Corticotropin-releasing factor	CRF	CRF	PF00473

*Highlighted **extracellular** domain families found to be multilocale.

Description	SMART name	Pfam name	Pfam ID
Granulocyte-macrophage colony-stimulating factor (GM-CSF)	CSF2	GM_CSF	PF01109
Domain first found in C1r, C1s, uEGF, and bone morphogenetic protein	CUB	CUB	PF00431
Defensin/corticostatin family	DEFSN	Defensin_beta	PF00711
Homologues of snake disintegrins	DISIN	Disintegrin	PF00200
DM4/DM12 family of domains in Drosophila melanogaster proteins of unknown function	DM4_12	DM4_12	PF07841
Cysteine-rich domain currently specific to Drosophila	DM6	DUF1431	PF07248
Possible catecholamine-binding domain present in a variety of eukaryotic proteins	DoH	DOMON	PF03351
Delta serrate ligand	DSL	DSL	PF01414
Epidermal growth factor-like domain	EGF	EGF	PF00008
Calcium-binding EGF-like domain	EGF_CA	EGF_CA	PF07645
Laminin-type epidermal growth factor-like domain	EGF_Lam	Laminin_EGF	PF00053
Endothelin	END	Endothelin	PF00322
Ependymins	EPEND	Ependymin	PF00811
Ephrin receptor ligand binding domain	EPH_lbd	Ephrin_lbd	PF01404
Four repeated domains in the Fasciclin I family of proteins, present in many other contexts	FAS1	Fasciclin	PF02469
Fibrinogen-related domains (FReDs)	FBG	Fibrinogen_C	PF00147
Fungal-type cellulose-binding domain	fCBD	CBM_1	PF00734
Fibronectin type 1 domain	FN1	fn1	PF00039
Fibronectin type 2 domain	FN2	fn2	PF00040
Frizzled	FRI	Fz	PF01392
Furin-like repeats	FU	Furin-like	PF00757
G2 nidogen domain and fibulin	G2F	G2F	PF07474
Gastrin / cholecystokinin / caerulein family	GASTRIN	Gastrin	PF00918
Glycoprotein hormone alpha chain homologues	GHA	Hormone_6	PF00236
Glycoprotein hormone beta chain homologues	GHB	Cys_knot	PF00007
Domain containing Gla (gamma-carboxyglutamate) residues	GLA	Gla	PF00594
Glucagon like hormones	GLUCA	Hormone_2	PF00123
Glycosyl hydrolase family 10	Glyco_10	Glyco_hydro_10	PF00331
Glyco_18	Glyco_18	Glyco_hydro_18	PF00704
Glycosyl hydrolases family 25	Glyco_25	Glyco_hydro_25	PF01183
Glycosyl hydrolases family 32	Glyco_32	Glyco_hydro_32	PF00251
G-protein-coupled receptor proteolytic site domain	GPS	GPS	PF01825
Granulin	GRAN	Granulin	PF00396
Hint (Hedgehog/Intein) domain C-terminal region	HintC	Hint	PF01079
Hint (Hedgehog/Intein) domain N-terminal region	HintN	Hint	PF01079

*Highlighted **extracellular** domain families found to be multilocal.

Description	SMART name	Pfam name	Pfam ID
Domain present in hormone receptors	HormR	HRM	PF02793
Hemopexin-like repeats	HX	Hemopexin	PF00045
Hydrophobins	HYDRO	Hydrophobin	PF01185
Insulin growth factor-binding protein homologues	IB	IGFBP	PF00219
Interferon alpha, beta and delta	IFabd	Interferon	PF00143
Interleukin-10 family	IL10	IL10	PF00726
Interleukin-2 family	IL2	IL2	PF00715
Interleukins 4 and 13	IL4_13	IL4	PF00727
Interleukin-6 homologues	IL6	IL6	PF00489
Interleukin-7 and interleukin-9 family	IL7	IL7	PF01415
Insulin / insulin-like growth factor / relaxin family	IIGF	Insulin	PF00049
Integrin beta subunits (N-terminal portion of extracellular region)	INB	Integrin_beta	PF00362
Integrin alpha (beta-propellor repeats)	Int_alpha	Integrin_alpha	PF00357
Kazal type serine protease inhibitors	KAZAL	Kazal_1	PF00050
Kazal type serine protease inhibitors	KAZAL	Kazal_2	PF07648
Knottins	Knot1	Defensin_2	PF01097
Kringle domain	KR	Kringle	PF00051
BPTI/Kunitz family of serine protease inhibitors	KU	Kunitz_BPTI	PF00014
Laminin B domain	LamB	Laminin_B	PF00052
Laminin G domain	LamG	Laminin_G_1	PF00054
Laminin N-terminal domain (domain VI)	LamNT	Laminin_N	PF00055
LCCL	LCCL	LCCL	PF03815
Low-density lipoprotein receptor domain class A	LDLa	Ldl_recept_a	PF00057
Leukemia inhibitory factor	LIF_OSM	LIF_OSM	PF01291
Link (Hyaluronan-binding)	LINK	Xlink	PF00193
Lipoprotein N-terminal Domain	LPD_N	Vitellogenin_N	PF01347
Ly-6 antigen / uPA receptor -like domain	LU	UPAR_LY6	PF00021
Low-density lipoprotein-receptor YWTD domain	LY	Ldl_recept_b	PF00058
Alpha-lactalbumin / lysozyme C	LYZ1	Lys	PF00062
Lysozyme subfamily 2	LYZ2	Glucosaminidase	PF01832
Membrane-attack complex / perforin	MACPF	MACPF	PF01823
Domain in meprin, A5, receptor protein tyrosine phosphatase mu (and others)	MAM	MAM	PF00629
Domain involved in innate immunity and lipid metabolism	ML	E1_DerP2_DerF2	PF02221
NEAr Transporter domain	NEAT	NEAT	PF05031
Nerve growth factor (NGF or beta-NGF)	NGF	NGF	PF00243
Neurohypophysial hormones	NH	Hormone_5	PF00184

*Highlighted **extracellular** domain families found to be multilocal.

Description	SMART name	Pfam name	Pfam ID
Extracellular domain of unknown function in nidogen (entactin) and hypothetical proteins	NIDO	NIDO	PF06119
Domain found in Notch and Lin-12	NL	Notch	PF00066
Neuromedin U	NMU	NMU	PF02070
Tissue inhibitor of metalloproteinase family	NTR	TIMP	PF00965
DNA/RNA non-specific endonuclease*	NUC	Endonuclease_NS	PF01223
Olfactomedin-like domains	OLF	OLF	PF02191
Osteopontin	OSTEO	Osteopontin	PF00865
Phospholipase A2	PA2c	Phospholip_A2_1	PF00068
Pancreatic hormones / neuropeptide F / peptide YY family	PAH	Hormone_3	PF00159
Divergent subfamily of APPLE domains	PAN_AP	PAN	PF00024
Bacterial periplasmic substrate-binding proteins	PBPb	SBP_bac_3	PF00497
P or trefoil or TFF domain	PD	Trefoil	PF00088
Platelet-derived and vascular endothelial growth factors (PDGF, VEGF) family	PDGF	PDGF	PF00341
Papain family cysteine protease	Pept_C1	Peptidase_C1	PF00112
Insect pheromone/odorant binding protein domains	PhBP	PBP_GOBP	PF01395
Repeats in polycystic kidney disease 1 (PKD1) and other proteins	PKD	PKD	PF00801
Major prion protein	PRP	Prion	PF00377
Paramecium Surface Antigen Repeat	PSA	Paramecium_SA	PF01508
Domain found in Plexins, Semaphorins and Integrins	PSI	PSI	PF01437
Plant trypsin inhibitors	PTI	Squash	PF00299
Pentraxin / C-reactive protein / pentaxin family	PTX	Pentaxin	PF00354
Pancreatic ribonuclease	RNase_Pc	RnaseA	PF00074
Serum amyloid A proteins	SAA	SAA	PF00277
Saposin/surfactant protein-B A-type DOMAIN	SAPA	SapA	PF02199
Saposin (B) Domains	SapB	SapB_1	PF05184
Saposin (B) Domains	SapB	SapB_2	PF03489
SCP / Tpx-1 / Ag5 / PR-1 / Sc7 family of extracellular domains	SCP	SCP	PF00188
Interocrine alpha family (small cytokine C-X-C) (chemokine CXC)	SCY	IL8	PF00048
Domain found in sea urchin sperm protein, enterokinase, agrin	SEA	SEA	PF01390
Semaphorin domain	Sema	Sema	PF01403
ShK toxin domain	ShKT	ShTK	PF01549
Somatomedin B -like domains	SO	Somatomedin_B	PF01033
Scavenger receptor Cys-rich	SR	SRCR	PF00530

*Highlighted **extracellular** domain families found to be multilocal.

Description	SMART name	Pfam name	Pfam ID
Soybean trypsin inhibitor (Kunitz) family of protease inhibitors	STI	Kunitz_legume	PF00197
Transforming growth factor-beta (TGF-beta) family	TGFB	TGF_beta	PF00019
Thaumatococcus family	THN	Thaumatococcus	PF00314
Tachykinin family	TK	Tachykinin	PF02202
Tumour necrosis factor family	TNF	TNF	PF00229
Tumor necrosis factor receptor / nerve growth factor receptor repeats	TNFR	TNFR_c6	PF00020
Transferrin	TR_FER	Transferrin	PF00405
Transthyretin	TR_THY	Transthyretin	PF00576
Trypsin-like serine protease	Tryp_SPc	Trypsin	PF00089
Thrombospondin type 1 repeats	TSP1	TSP_1	PF00090
Thyroglobulin type I repeats	TY	Thyroglobulin_1	PF00086
Uteroglobin	UTG	Uteroglobin	PF01099
von Willebrand factor (vWF) type C domain	VWC	VWC	PF00093
von Willebrand factor (vWF) type D domain	VWD	VWD	PF00094
Four-disulfide core domains	WAP	WAP	PF00095
Wnt-inhibitory factor-1 like domain	WIF	WIF	PF02019
Found in Wnt-1	WNT1	wnt	PF00110
Present in yeast cell wall integrity and stress response component proteins	WSC	WSC	PF01822
Zinc-dependent metalloprotease	ZnMc	Peptidase_M10	PF00413
Zona pellucida (ZP) domain	ZP	Zona_pellucida	PF00100
ADAM-TS Spacer 1	-	ADAM_spacer1	PF05986
Animal haem peroxidase	-	An_peroxidase	PF03098
Antistatin	-	Antistatin	PF02822
Copper type II ascorbate-dependent monooxygenase, C-terminal domain	-	Cu2_monoox_C	PF03712
Copper type II ascorbate-dependent monooxygenase, N-terminal domain	-	Cu2_monooxygen	PF01082
Laminin Domain I	-	Laminin_I	PF06008
Laminin Domain II	-	Laminin_II	PF06009
Laminin G domain	TSPN	Laminin_G_2	PF02210
Class I Histocompatibility antigen, domains alpha 1 and 2	-	MHC_I	PF00129
Class II histocompatibility antigen, beta domain	-	MHC_II_beta	PF00969
Agrin NtA domain	-	NtA	PF03146
Reprolysin (M12B) family zinc metalloprotease	-	Reprolysin	PF01421

*Highlighted **extracellular** domain families found to be multilocal.

Additional file 4. List of intracellular Pfam-A signaling domain families. The table contains obligatory cytoplasmic Pfam-A domain families which were used to predict subcellular localization of proteins. Our domain co-occurrence analyses of Metazoan UniProtKB entries have identified 115 obligatory cytoplasmic Pfam-A domain families, the majority of which are also identified as such in the SMART database.

Description	SMART name	Pfam name	Pfam ID
14-3-3 homologues	14_3_3	14-3-3	PF00244
Actin depolymerisation factor/cofilin -like domains	ADF	Cofilin_ADF	PF00241
Ankyrin repeats	ANK	Ank	PF00023
ARF-like small GTPases; ARF, ADP-ribosylation factor	ARF	Arf	PF00025
Putative GTP-ase activating proteins for the small GTPase, ARF	ArfGap	ArfGap	PF01412
Band 4.1 homologues	B41	Band_41	PF00373
Cytochrome b-561 / ferric reductase transmembrane domain	B561	Cytochrom_B561	PF03188
Bulb-type mannose-specific lectin	B_lectin	B_lectin	PF01453
BAG domains, present in regulator of Hsp70 proteins	BAG	BAG	PF02179
BH4 Bcl-2 homology region 4	BH4	BH4	PF02180
Bruton's tyrosine kinase Cys-rich motif	BTK	BTK	PF00779
Protein kinase C conserved region 1 (C1) domains (Cysteine-rich domains)	DAG_PE-bind	C1_1	PF00130
Calponin homology domain	CH	CH	PF00307
Two component signalling adaptor domain	CheW	CheW	PF01584
Domain found in NIK1-like kinases, mouse citron and yeast ROM1, ROM2	CNH	CNH	PF00780
Cyclic nucleotide-monophosphate binding domain	cNMP	cNMP_binding	PF00027
Cullin	CULLIN	Cullin	PF00888
Adenylyl- / guanylyl cyclase, catalytic domain	CYCc	Guanylate_cyc	PF00211
Diacylglycerol kinase accessory domain (presumed)	DAGKa	DAGK_acc	PF00609
Diacylglycerol kinase catalytic domain (presumed)	DAGKc	DAGK_cat	PF00781
Domain present in Dishevelled and axin	DAX	DIX	PF00778
Death effector domain	DED	DED	PF01335
Domain found in Dishevelled, Egl-10, and Pleckstrin	DEP	DEP	PF00610
Dual specificity phosphatase, catalytic domain	DSPc	DSPc	PF00782
Domain of Unknown Function with GGDEF motif	DUF1	GGDEF	PF00990
Domain of Unknown Function 2	DUF2	EAL	PF00563
Dynamin, GTPase	DYNc	Dynamin_N	PF00350
Epsin N-terminal homology (ENTH) domain	ENTH	ENTH	PF01417
A Receptor for Ubiquitination Targets	FBOX	F-box	PF00646
Fes/CIP4 homology domain	FCH	FCH	PF00611
Contains two conserved F residues	FF	FF	PF01846
Formin Homology 2 Domain	FH2	FH2	PF02181
Forkhead associated domain	FHA	FHA	PF00498
Protein present in Fab1, YOTB, Vac1, and EEA1	FYVE	FYVE	PF01363
Growth-Arrest-Specific Protein 2 Domain	GAS2	GAS2	PF02187
Dynamin GTPase effector domain	GED	GED	PF02212

Description	SMART name	Pfam name	Pfam ID
G protein gamma subunit-like motifs	GGL	G-gamma	PF00631
LGN motif, putative GEFs specific for G-alpha GTPases	GoLoco	GoLoco	PF02188
Domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins	GRAM	GRAM	PF02893
Guanylate kinase homologues	GuKc	Guanylate_kin	PF00625
HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases) domain	HAMP	HAMP	PF00672
Histidine kinase-like ATPases	HATPase_c	HATPase_c	PF02518
Domain Homologous to E6-AP Carboxyl Terminus with	HECTc	HECT	PF00632
His Kinase A (phosphoacceptor) domain	HisKA	HisKA	PF00512
Histidine Phosphotransfer domain	HPT	Hpt	PF01627
Rho effector or protein kinase C-related kinase homology region 1 homologues	Hr1	HR1	PF02185
I/LWEQ domain	ILWEQ	I_LWEQ	PF01608
Inositol polyphosphate phosphatase, catalytic domain homologues	IPPc	Exo_endo_phos	PF03372
Short calmodulin-binding motif containing conserved Ile and Gln residues	IQ	IQ	PF00612
Immunoreceptor tyrosine-based activation motif	ITAM	ITAM	PF02189
Kinesin motor, catalytic domain. ATPase	KISc	Kinesin	PF00225
Domain in receptor targeting proteins Lin-2 and Lin-7	L27	L27	PF02828
Zinc-binding domain present in Lin-11, Isl-1, Mec-3	LIM	LIM	PF00412
Low molecular weight phosphatase family	LMWPc	LMWPc	PF01451
Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer)	MA	MCPsignal	PF00015
Myosin	MYSc	Myosin_head	PF00063
Domain in Myosin and Kinesin Tails	MyTH4	MyTH4	PF00784
Motif C-terminal to PAS motifs (likely to contribute to PAS structural domain)	PAC	PAC	PF00785
PAS domain	PAS	PAS	PF00989
PASTA	PASTA	PASTA	PF03793
P21-Rho-binding domain	PBD	PBD	PF00786
Pleckstrin homology domain	PH	PH	PF00169
Phosphoinositide 3-kinase, region postulated to contain C2 domain	PI3K_C2	PI3K_C2	PF00792
PI3-kinase family, p85-binding domain	PI3K_p85B	PI3K_p85B	PF02192
PI3-kinase family, Ras-binding domain	PI3K_rbd	PI3K_rbd	PF00794
Phosphoinositide 3-kinase family, accessory domain (PIK domain)	PI3Ka	PI3Ka	PF00613
Phosphoinositide 3-kinase, catalytic domain	PI3Kc	PI3_PI4_kinase	PF00454
Phosphatidylinositol phosphate kinases	PIPKc	PIP5K	PF01504
Cytoplasmic phospholipase A2, catalytic subunit	PLAc	PLA2_B	PF01735

Description	SMART name	Pfam name	Pfam ID
Phospholipase C, catalytic domain (part); domain X	PLCXc	PI-PLC-X	PF00388
Phospholipase C, catalytic domain (part); domain Y	PLCYc	PI-PLC-Y	PF00387
Phospholipase D. Active site motifs	PLDc	PLDc	PF00614
Serine/threonine phosphatases, family 2C, catalytic domain	PP2Cc	PP2C	PF00481
Phosphotyrosine-binding domain, phosphotyrosine-interaction (PI) domain	PTB	PID	PF00640
Phosphotyrosine-binding domain (IRS1-like)	PTBI	IRS	PF02174
Protein tyrosine phosphatase, catalytic domain	PTPc	Y_phosphatase	PF00102
Protein tyrosine phosphatase, catalytic domain, undefined specificity	PTPc_DSPc	Y_phosphatase	PF00102
Protein tyrosine phosphatase, catalytic domain motif	PTPc_motif	Y_phosphatase	PF00102
PhoX homologous domain, present in p47phox and p40phox	PX	PX	PF00787
Domain associated with PX domains	PXA	PXA	PF02194
Ras association (RalGDS/AF-6) domain	RA	RA	PF00788
Ran-binding domain	RanBD	Ran_BP1	PF00638
Ras subfamily of RAS small GTPases	RAS	Ras	PF00071
GTPase-activator protein for Ras-like GTPases	RasGAP	RasGAP	PF00616
Guanine nucleotide exchange factor for Ras-like small GTPases	RasGEF	RasGEF	PF00617
Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	RasGEFN	RasGEF_N	PF00618
Raf-like Ras-binding domain	RBD	RBD	PF02196
Regulator of G protein signalling domain	RGS	RGS	PF00615
GTPase-activator protein for Rho-like GTPases	RhoGAP	RhoGAP	PF00620
Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases	RhoGEF	RhoGEF	PF00621
RIIalpha, Regulatory subunit portion of type II PKA R-subunit	RIIa	RIIa	PF02197
RIO-like kinase	RIO	RIO1	PF01163
Extension to Ser/Thr-type protein kinases	S_TK_X	Pkinase_C	PF00433
Serine/Threonine protein kinases, catalytic domain	S_TKc	Pkinase	PF00069
Sterile alpha motif	SAM	SAM_1	PF00536
Sterile alpha motif	SAM	SAM_2	PF07647
Sec7 domain	Sec7	Sec7	PF01369
Src homology 2 domains	SH2	SH2	PF00017
Suppressors of cytokine signalling	SOCS	SOCS_box	PF07525
Spectrin repeats	SPEC	Spectrin	PF00435
In StAR and phosphatidylcholine transfer protein	START	START	PF01852
Protein kinase; unclassified specificity	STYKc	Pkinase	PF00069
Helical region found in SNAREs	t_SNARE	SNARE	PF05739

Description	SMART name	Pfam name	Pfam ID
Homologues of the ligand binding domain of Tar	TarH	TarH	PF02203
Domain in Tre-2, BUB2p, and Cdc16p. Probable Rab-GAPs	TBC	TBC	PF00566
Metallochaperone-like domain	TRASH	YHS	PF04945
Tyrosine kinase, catalytic domain	TyrKc	Pkinase	PF00069
Ubiquitin associated domain	UBA	UBA	PF00627
Ubiquitin homologues	UBQ	ubiquitin	PF00240
Domain present in ubiquitin-regulatory proteins	UBX	UBX	PF00789
Domain present in VPS-27, Hrs and STAM	VHS	VHS	PF00790
Domain present in VPS9	VPS9	VPS9	PF02204
WD40 repeats	WD40	WD40	PF00400
WASP homology region 1	WH1	WH1	PF00568
Wiskott Aldrich syndrome homology region 2	WH2	WH2	PF02205
Domain with 2 conserved Trp (W) residues	WW	WW	PF00397
A20-like zinc fingers	ZnF_A20	zf-A20	PF01754
Zinc-binding domain, present in Dystrophin, CREB-binding protein	ZnF_ZZ	ZZ	PF00569
Domain present in ZO-1 and Unc5-like netrin receptors	ZU5	ZU5	PF00791

Additional file 5. List of nuclear Pfam-A domain families. The table contains obligatory nuclear Pfam-A domain families which were used to predict subcellular localization of proteins. Our domain co-occurrence analyses of Metazoan UniProtKB entries have identified 126 obligatory nuclear Pfam-A domain families, the majority of which are also identified as such in the SMART database.

Description	SMART name	Pfam name	Pfam ID
3'-5' exonuclease	35EXOc	3_5_exonuc	PF01612
5'-3' exonuclease	53EXOc	5_3_exonuc	PF01367
tRNA-specific and double-stranded RNA adenosine deaminase (RNA-specific editase)	ADEAMc	A_deamin	PF02137
DNA-binding domain in plant proteins such as APETALA2 and EREBPs	AP2	AP2	PF00847
AP endonuclease family 2	AP2Ec	AP_endonuc_2	PF01261
Bromo adjacent homology domain	BAH	BAH	PF01426
Basic domain in HLH proteins of MYOD family	BASIC	Basic	PF01586
B-Box-type zinc finger	BBOX	zf-B_box	PF00643
BCL (B-Cell lymphoma); contains BH1, BH2 regions	BCL	Bcl-2	PF00452
Breast cancer carboxy-terminal domain	BRCT	BRCT	PF00533
BRIGHT, ARID (A/T-rich interaction domain) domain	BRIGHT	ARID	PF01388
Domain in transcription and CHROMO domain helicases	BRK	TCH	PF07533
Basic region leucin zipper	BRLZ	bZIP	PF00170
Bromo domain	BROMO	Bromodomain	PF00439
Domain in transcription factors and synapse-associated proteins	BSD	BSD	PF03909
Bromodomain transcription factors and PHD domain containing proteins	BTP	Bromo_TP	PF07524
CCAAT-Binding transcription Factor	CBF	CBFB_NFYA	PF02045
Chromatin organization modifier domain	CHROMO	Chromo	PF00385
Catalytic domain of ctd-like phosphatases	CPDc	NIF	PF03031
Cold shock protein domain	CSP	CSD	PF00313
Copper-Fist	Cu_Fist	Copper-fist	PF00649
Domain in different transcription and chromosome remodeling factors	DDT	DDT	PF02791
DEAD-like helicases superfamily	DEXDc	SNF2_N	PF00176
Doublesex DNA-binding motif	DM	DM	PF00751
Double-stranded RNA binding motif	DSRM	dsrm	PF00035
Domain A in dwarfin family proteins	DWA	MH1	PF03165
Domain B in dwarfin family proteins	DWB	MH2	PF03166
Domain in DSRM or ZnF_C2H2 domain containing proteins	DZF	DZF	PF07528
Domain at the C-termini of GCD6, eIF-2B epsilon, eIF-4 gamma and eIF-5	eIF5C	W2	PF02020
Elongator protein 3, MiaB family, Radical SAM	Elp3	Radical_SAM	PF04055
Endonuclease III	ENDO3c	HhH-GPD	PF00730
Erythroblast transformation specific domain	ETS	Ets	PF00178

Description	SMART name	Pfam name	Pfam ID
EXOIII	EXOIII	Exonuc_X-T	PF00929
FORKHEAD	FH	Fork_head	PF00250
"FY-rich" domain, C-terminal region	FYRC	FYRC	PF05965
"FY-rich" domain, N-terminal region	FYRN	FYRN	PF05964
Glycine rich nucleic binding domain	G_patch	G-patch	PF01585
GIY-YIG type nucleases (URI domain)	GIYc	GIY-YIG	PF01541
Domain in histone families 1 and 5	H15	Linker_histone	PF00538
Histone 2A	H2A	Histone	PF00125
Histone H2B	H2B	Histone	PF00125
Histone H3	H3	Histone	PF00125
Histone H4	H4	Histone	PF00125
Homeobox associated leucin zipper	HALZ	HALZ	PF02183
Helicase superfamily c-terminal domain	HELICc	Helicase_C	PF00271
Helix-hairpin-helix DNA-binding motif class 1	HhH1	HHH	PF00633
High mobility group	HMG	HMG_box	PF00505
Domain in high mobility group proteins HMG14 and HMG 17	HMG17	HMG14_17	PF01101
HNH nucleases	HNHc	HNH	PF01844
Domain in histone-like proteins of HNS family	HNS	Histone_HNS	PF00816
Ligand binding domain of hormone receptors	HOLI	Hormone_recep	PF00104
Helicase and RNase D C-terminal	HRDC	HRDC	PF00570
Domain in helicases and associated with SANT domains	HSA	HSA	PF07529
Helix_turn_helix, arabinose operon control protein	HTH_ARAC	HTH_AraC	PF00165
Helix_turn_helix, Arsenical Resistance Operon Repressor	HTH_ARSR	HTH_5	PF01022
Helix_turn_helix ASNC type	HTH_ASNC	AsnC_trans_reg	PF01037
Helix_turn_helix, Deoxyribose operon repressor	HTH_DEOR	DeoR	PF00455
Helix-turn-helix diptheria tox regulatory element	HTH_DTXR	Fe_dep_repress	PF01325
Helix_turn_helix gluconate operon transcriptional repressor	HTH_GNTR	GntR	PF00392
Helix_turn_helix lactose operon repressor	HTH_LACI	LacI	PF00356
Helix_turn_helix multiple antibiotic resistance protein	HTH_MARR	MarR	PF01047
Helix_turn_helix, mercury resistance	HTH_MERR	MerR	PF00376
Helix-turn-helix XRE-family like proteins	HTH_XRE	HTH_3	PF01381
Intron encoded nuclease repeat motif	IENR1	NUMOD1	PF07453
Intron-encoded nuclease repeat 2	IENR2	NUMOD3	PF07460
Interferon regulatory factor	IRF	IRF	PF00605
A domain family that is part of the cupin metalloenzyme superfamily	JmjC	JmjC	PF02373

Description	SMART name	Pfam name	Pfam ID
Small domain found in the jumonji family of transcription factors	JmjN	JmjN	PF02375
KOW (Kyprides, Ouzounis, Woese) motif	KOW	KOW	PF00467
In nuclear membrane-associated proteins	LEM	LEM	PF03020
Ligase N family	LIGANc	DNA_ligase_aden	PF01653
Domain in DAP-5, eIF4G, MA-3 and other proteins	MA3	MA3	PF02847
MADS	MADS	SRF-TF	PF00319
Methyl-CpG binding domain	MBD	MBD	PF01429
Minichromosome maintenance proteins	MCM	MCM	PF00493
ATPase domain of DNA mismatch repair MUTS family	MUTSac	MutS_V	PF00488
NEUZ	NEUZ	Neuralized	PF07177
Orange domain	ORANGE	Hairy_orange	PF07527
ParB-like nuclease domain	ParB	ParBc	PF02195
Paired Box domain	PAX	PAX	PF00292
PHD zinc finger	PHD	PHD	PF00628
Short conserved domain in transcriptional regulators	Plus3	Plus-3	PF03126
DNA polymerase III beta subunit	POL3Bc	DNA_pol3_beta	PF00712
DNA polymerase A domain	POLAc	DNA_pol_A	PF00476
DNA polymerase type-B family	POLBc	DNA_pol_B	PF00136
DNA polymerase alpha chain like domain	POLIIIAc	PHP	PF02811
A domain found in a protein subunit of human RNase MRP and RNase P ribonucleoprotein complexes and archaeal proteins	POP4	UPF0086	PF01868
Found in Pit-Oct-Unc transcription factors	POU	Pou	PF00157
Domain associated with HOX domains	POX	POX	PF07526
PRE_C2HC	PRE_C2HC	PRE_C2HC	PF07530
N-terminal to some SET domains	PreSET	Pre-SET	PF05033
Proline-rich domain in spliceosome associated proteins	PSP	PSP	PF04046
DNA/RNA-binding repeats in PUR-alpha/beta/gamma and in hypothetical proteins from spirochetes and the Bacteroides-Cytophaga-Flexibacter bacteria	PUR	PurA	PF04845
Ribonuclease III family	RIBOc	Ribonuclease_3	PF00636
DNA-directed RNA-polymerase II subunit	RPOL4c	RNA_pol_Rpb4	PF03874
RNA polymerase subunit 8	RPOL8c	RNA_pol_Rpb8	PF03870
RNA polymerase subunit 9	RPOL9	RNA_POL_M_15KD	PF02150
RNA polymerase I subunit A N-terminus	RPOLA_N	RNA_pol_Rpb1_2	PF00623
RNA polymerase subunit CX	RPOLCX	DNA_RNApol_7kD	PF03604
RNA polymerases D	RPOLD	RNA_pol_A_bac	PF01000

Description	SMART name	Pfam name	Pfam ID
RPR	RPR	DUF618	PF04818
S4 RNA-binding domain	S4	S4	PF01479
Putative DNA-binding (bihelical) motif predicted to be involved in chromosomal organisation	SAP	SAP	PF02037
SET (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain	SET	SET	PF00856
Found in Skp1 protein family	Skp1	Skp1	PF01466
snRNP Sm proteins	Sm	LSM	PF01423
SPK	SPK	SPK	PF04435
SET and RING finger associated domain	SRA	YDG_SRA	PF02182
STE like transcription factors	STE	STE	PF02200
Domain first found in the mice T locus (Brachyury) protein	TBOX	T-box	PF00907
TEA domain	TEA	TEA	PF01285
Transcription initiation factor IIE	TFIIE	TFIIE_alpha	PF02002
Domain in the central regions of transcription elongation factor S-II (and elsewhere)	TFS2M	TFIIS_M	PF07500
Bacterial DNA topoisomerase I DNA-binding domain	TOP1Ac	Topoisom_bac	PF01131
Bacterial DNA topoisomeraes I ATP-binding domain	TOP1Bc	Topoisom_bac	PF01131
DNA Topoisomerase I (eukaryota)	TOPEUc	Topoisom_I	PF01028
TOPRIM	TOPRIM	Toprim	PF01751
Xeroderma pigmentosum G I-region	XPGI	XPG_I	PF00867
Xeroderma pigmentosum G N-region	XPGN	XPG_N	PF00752
Z-DNA-binding domain in adenosine deaminases	Zalpha	z-alpha	PF02295
BED zinc finger	ZnF_BED	zf-BED	PF02892
C2C2 Zinc finger	ZnF_C2C2	TFIIS_C	PF01096
Zinc finger	ZnF_C2H2	zf-C2H2	PF00096
Zinc finger	ZnF_C2HC	zf-CCHC	PF00098
Zinc finger	ZnF_C3H1	zf-CCCH	PF00642
C4 zinc finger in nuclear hormone receptors	ZnF_C4	zf-C4	PF00105
Zinc finger	ZnF_CHCC	zf-CHC2	PF01807
Zinc finger in DBF-like proteins	ZnF_DBF	zf-DBF	PF07535
Zinc finger binding to DNA consensus sequence [AT]GATA[AG]	ZnF_GATA	GATA	PF00320
Plant mutator transposase zinc finger	ZnF_PMZ	SWIM	PF04434

Additional file 6. List of Pfam-A domain families suitable for the study of domain integrity. The table contains Pfam-A domain families present in human, vertebrate and metazoa+fungi Swiss-Prot proteins that have a well-defined, conserved sequence length range and thus proved to be suitable for the study of domain integrity. The detailed description of the identification of Pfam-A families suitable for the study of the domain integrity is found in Additional file 1, Part 2.

Human Pfam ID												
PF00001	PF00323	PF00694	PF01119	PF01549	PF02171	PF02823	PF03487	PF04263	PF05060	PF05997	PF07169	PF08202
PF00005	PF00324	PF00696	PF01120	PF01551	PF02172	PF02824	PF03489	PF04265	PF05063	PF06001	PF07175	PF08204
PF00006	PF00326	PF00698	PF01121	PF01553	PF02173	PF02825	PF03491	PF04272	PF05064	PF06008	PF07177	PF08210
PF00009	PF00327	PF00704	PF01122	PF01556	PF02174	PF02826	PF03493	PF04275	PF05071	PF06009	PF07189	PF08213
PF00010	PF00328	PF00705	PF01124	PF01557	PF02177	PF02827	PF03494	PF04280	PF05076	PF06010	PF07202	PF08215
PF00013	PF00329	PF00708	PF01125	PF01562	PF02178	PF02828	PF03501	PF04300	PF05083	PF06012	PF07221	PF08216
PF00014	PF00330	PF00709	PF01126	PF01564	PF02179	PF02836	PF03508	PF04376	PF05089	PF06017	PF07225	PF08221
PF00017	PF00333	PF00711	PF01127	PF01565	PF02180	PF02837	PF03509	PF04377	PF05090	PF06025	PF07260	PF08232
PF00019	PF00334	PF00714	PF01129	PF01566	PF02181	PF02838	PF03511	PF04379	PF05091	PF06026	PF07263	PF08235
PF00021	PF00335	PF00715	PF01130	PF01569	PF02182	PF02840	PF03516	PF04382	PF05093	PF06031	PF07264	PF08240
PF00024	PF00337	PF00717	PF01131	PF01571	PF02184	PF02841	PF03517	PF04388	PF05104	PF06046	PF07281	PF08241
PF00025	PF00338	PF00719	PF01133	PF01575	PF02185	PF02843	PF03520	PF04389	PF05110	PF06049	PF07286	PF08242
PF00026	PF00339	PF00727	PF01134	PF01582	PF02186	PF02844	PF03521	PF04406	PF05111	PF06052	PF07292	PF08246
PF00027	PF00341	PF00728	PF01135	PF01583	PF02187	PF02845	PF03522	PF04408	PF05118	PF06058	PF07297	PF08264
PF00029	PF00342	PF00731	PF01138	PF01585	PF02189	PF02847	PF03523	PF04410	PF05127	PF06060	PF07303	PF08265
PF00030	PF00343	PF00732	PF01140	PF01586	PF02190	PF02852	PF03528	PF04419	PF05131	PF06068	PF07324	PF08266
PF00032	PF00344	PF00733	PF01142	PF01588	PF02191	PF02854	PF03529	PF04420	PF05132	PF06071	PF07334	PF08271
PF00033	PF00346	PF00735	PF01144	PF01590	PF02192	PF02862	PF03530	PF04421	PF05147	PF06079	PF07347	PF08279
PF00034	PF00347	PF00736	PF01146	PF01591	PF02194	PF02864	PF03531	PF04423	PF05148	PF06083	PF07353	PF08285
PF00037	PF00348	PF00743	PF01148	PF01592	PF02196	PF02865	PF03533	PF04427	PF05153	PF06087	PF07354	PF08288
PF00038	PF00349	PF00748	PF01150	PF01596	PF02197	PF02866	PF03535	PF04433	PF05158	PF06105	PF07359	PF08292
PF00039	PF00350	PF00749	PF01151	PF01597	PF02198	PF02867	PF03540	PF04434	PF05160	PF06119	PF07400	PF08295
PF00040	PF00351	PF00750	PF01153	PF01598	PF02199	PF02870	PF03542	PF04437	PF05162	PF06121	PF07406	PF08311
PF00042	PF00352	PF00751	PF01154	PF01599	PF02201	PF02872	PF03546	PF04438	PF05181	PF06140	PF07412	PF08313

PF00044	PF00354	PF00752	PF01157	PF01602	PF02202	PF02874	PF03547	PF04440	PF05184	PF06148	PF07421	PF08318
PF00045	PF00355	PF00753	PF01158	PF01603	PF02204	PF02875	PF03556	PF04442	PF05185	PF06179	PF07443	PF08321
PF00046	PF00357	PF00755	PF01159	PF01608	PF02205	PF02877	PF03561	PF04487	PF05186	PF06189	PF07448	PF08324
PF00048	PF00361	PF00756	PF01160	PF01619	PF02207	PF02879	PF03567	PF04488	PF05187	PF06201	PF07452	PF08326
PF00049	PF00362	PF00757	PF01161	PF01624	PF02208	PF02880	PF03568	PF04494	PF05188	PF06202	PF07458	PF08327
PF00050	PF00363	PF00758	PF01162	PF01625	PF02209	PF02881	PF03571	PF04495	PF05189	PF06209	PF07469	PF08332
PF00051	PF00364	PF00762	PF01163	PF01630	PF02212	PF02882	PF03572	PF04503	PF05190	PF06211	PF07474	PF08333
PF00052	PF00365	PF00763	PF01166	PF01633	PF02213	PF02883	PF03577	PF04505	PF05191	PF06212	PF07479	PF08336
PF00054	PF00366	PF00764	PF01167	PF01642	PF02214	PF02885	PF03587	PF04511	PF05193	PF06214	PF07496	PF08337
PF00055	PF00368	PF00766	PF01168	PF01652	PF02218	PF02887	PF03600	PF04515	PF05195	PF06218	PF07500	PF08344
PF00056	PF00370	PF00769	PF01170	PF01655	PF02219	PF02888	PF03604	PF04538	PF05196	PF06220	PF07502	PF08351
PF00058	PF00375	PF00773	PF01172	PF01661	PF02221	PF02889	PF03607	PF04547	PF05199	PF06221	PF07521	PF08357
PF00060	PF00377	PF00774	PF01175	PF01663	PF02223	PF02891	PF03615	PF04548	PF05207	PF06229	PF07522	PF08365
PF00062	PF00378	PF00777	PF01176	PF01667	PF02225	PF02892	PF03619	PF04549	PF05208	PF06237	PF07524	PF08368
PF00066	PF00380	PF00778	PF01179	PF01669	PF02229	PF02893	PF03623	PF04553	PF05222	PF06248	PF07525	PF08373
PF00068	PF00382	PF00779	PF01180	PF01687	PF02230	PF02897	PF03630	PF04557	PF05236	PF06268	PF07527	PF08374
PF00072	PF00383	PF00780	PF01182	PF01694	PF02233	PF02898	PF03635	PF04558	PF05237	PF06292	PF07528	PF08375
PF00074	PF00384	PF00781	PF01186	PF01699	PF02234	PF02902	PF03637	PF04560	PF05240	PF06294	PF07529	PF08377
PF00075	PF00385	PF00782	PF01187	PF01702	PF02237	PF02910	PF03643	PF04561	PF05241	PF06297	PF07531	PF08383
PF00077	PF00386	PF00784	PF01189	PF01704	PF02238	PF02911	PF03645	PF04565	PF05250	PF06309	PF07533	PF08384
PF00080	PF00387	PF00786	PF01191	PF01709	PF02244	PF02913	PF03647	PF04566	PF05251	PF06311	PF07539	PF08385
PF00081	PF00388	PF00787	PF01192	PF01712	PF02245	PF02919	PF03651	PF04567	PF05253	PF06312	PF07540	PF08389
PF00082	PF00390	PF00789	PF01194	PF01713	PF02251	PF02920	PF03660	PF04568	PF05254	PF06320	PF07541	PF08390
PF00083	PF00393	PF00790	PF01198	PF01715	PF02252	PF02921	PF03661	PF04571	PF05255	PF06327	PF07542	PF08391
PF00084	PF00394	PF00791	PF01199	PF01722	PF02257	PF02922	PF03665	PF04572	PF05276	PF06328	PF07545	PF08393
PF00085	PF00396	PF00792	PF01200	PF01725	PF02259	PF02928	PF03666	PF04573	PF05279	PF06331	PF07546	PF08397
PF00088	PF00397	PF00794	PF01201	PF01729	PF02260	PF02931	PF03669	PF04579	PF05281	PF06333	PF07562	PF08399
PF00090	PF00398	PF00795	PF01202	PF01731	PF02262	PF02932	PF03670	PF04587	PF05282	PF06337	PF07565	PF08403
PF00091	PF00401	PF00797	PF01204	PF01733	PF02267	PF02933	PF03671	PF04588	PF05283	PF06345	PF07569	PF08412
PF00092	PF00402	PF00801	PF01207	PF01734	PF02268	PF02934	PF03676	PF04589	PF05287	PF06350	PF07571	PF08418
PF00093	PF00403	PF00804	PF01208	PF01735	PF02269	PF02935	PF03690	PF04592	PF05291	PF06365	PF07576	PF08420
PF00094	PF00405	PF00805	PF01210	PF01740	PF02270	PF02936	PF03694	PF04593	PF05292	PF06367	PF07593	PF08423

PF00095	PF00408	PF00806	PF01213	PF01743	PF02271	PF02937	PF03700	PF04597	PF05296	PF06372	PF07647	PF08424
PF00096	PF00410	PF00808	PF01214	PF01747	PF02272	PF02939	PF03712	PF04598	PF05300	PF06373	PF07648	PF08429
PF00098	PF00411	PF00809	PF01215	PF01749	PF02274	PF02946	PF03715	PF04614	PF05303	PF06374	PF07649	PF08430
PF00105	PF00412	PF00810	PF01216	PF01751	PF02275	PF02947	PF03719	PF04615	PF05324	PF06375	PF07650	PF08433
PF00106	PF00414	PF00812	PF01217	PF01753	PF02284	PF02953	PF03720	PF04617	PF05326	PF06384	PF07651	PF08441
PF00107	PF00416	PF00814	PF01218	PF01754	PF02285	PF02961	PF03721	PF04621	PF05328	PF06387	PF07657	PF08442
PF00108	PF00418	PF00817	PF01221	PF01756	PF02290	PF02965	PF03725	PF04627	PF05337	PF06391	PF07662	PF08447
PF00109	PF00420	PF00821	PF01222	PF01758	PF02291	PF02966	PF03726	PF04628	PF05347	PF06393	PF07670	PF08449
PF00111	PF00428	PF00822	PF01223	PF01759	PF02296	PF02969	PF03727	PF04629	PF05349	PF06396	PF07677	PF08450
PF00112	PF00433	PF00826	PF01227	PF01762	PF02301	PF02970	PF03730	PF04636	PF05350	PF06399	PF07678	PF08451
PF00113	PF00436	PF00827	PF01229	PF01765	PF02310	PF02971	PF03731	PF04641	PF05351	PF06400	PF07684	PF08454
PF00115	PF00439	PF00828	PF01230	PF01770	PF02312	PF02978	PF03735	PF04643	PF05355	PF06401	PF07687	PF08466
PF00116	PF00441	PF00831	PF01231	PF01773	PF02318	PF02984	PF03736	PF04644	PF05361	PF06418	PF07699	PF08473
PF00117	PF00444	PF00832	PF01233	PF01775	PF02319	PF02985	PF03747	PF04652	PF05362	PF06419	PF07700	PF08474
PF00118	PF00445	PF00833	PF01234	PF01776	PF02320	PF02990	PF03758	PF04658	PF05365	PF06424	PF07701	PF08477
PF00119	PF00446	PF00834	PF01237	PF01777	PF02330	PF02991	PF03764	PF04664	PF05366	PF06427	PF07703	PF08487
PF00120	PF00447	PF00835	PF01238	PF01778	PF02337	PF02996	PF03765	PF04667	PF05378	PF06428	PF07706	PF08490
PF00121	PF00448	PF00836	PF01239	PF01779	PF02338	PF03002	PF03770	PF04675	PF05383	PF06432	PF07710	PF08491
PF00123	PF00450	PF00837	PF01242	PF01780	PF02344	PF03006	PF03781	PF04679	PF05386	PF06441	PF07711	PF08492
PF00125	PF00452	PF00838	PF01243	PF01781	PF02345	PF03009	PF03792	PF04680	PF05392	PF06444	PF07716	PF08499
PF00129	PF00454	PF00839	PF01244	PF01784	PF02347	PF03016	PF03798	PF04683	PF05395	PF06446	PF07717	PF08504
PF00130	PF00458	PF00850	PF01245	PF01791	PF02348	PF03020	PF03803	PF04691	PF05399	PF06448	PF07718	PF08506
PF00131	PF00459	PF00852	PF01246	PF01794	PF02350	PF03024	PF03807	PF04692	PF05404	PF06455	PF07719	PF08510
PF00132	PF00462	PF00853	PF01247	PF01798	PF02351	PF03028	PF03810	PF04695	PF05405	PF06456	PF07722	PF08513
PF00133	PF00464	PF00855	PF01248	PF01799	PF02354	PF03029	PF03813	PF04696	PF05406	PF06459	PF07723	PF08514
PF00135	PF00466	PF00856	PF01249	PF01803	PF02359	PF03031	PF03815	PF04697	PF05422	PF06461	PF07724	PF08515
PF00136	PF00467	PF00858	PF01251	PF01805	PF02366	PF03034	PF03820	PF04698	PF05427	PF06462	PF07731	PF08516
PF00137	PF00468	PF00860	PF01253	PF01808	PF02372	PF03036	PF03821	PF04699	PF05428	PF06463	PF07732	PF08517
PF00143	PF00472	PF00861	PF01254	PF01812	PF02373	PF03039	PF03823	PF04704	PF05434	PF06464	PF07738	PF08518
PF00144	PF00473	PF00864	PF01255	PF01813	PF02374	PF03045	PF03826	PF04706	PF05438	PF06465	PF07741	PF08519
PF00145	PF00474	PF00865	PF01257	PF01821	PF02375	PF03051	PF03827	PF04707	PF05439	PF06466	PF07742	PF08523
PF00146	PF00476	PF00867	PF01259	PF01822	PF02376	PF03054	PF03828	PF04709	PF05450	PF06467	PF07743	PF08526

PF00149	PF00479	PF00868	PF01262	PF01823	PF02377	PF03055	PF03832	PF04710	PF05454	PF06468	PF07744	PF08527
PF00151	PF00480	PF00870	PF01263	PF01826	PF02383	PF03056	PF03834	PF04711	PF05456	PF06469	PF07748	PF08534
PF00152	PF00481	PF00879	PF01265	PF01833	PF02389	PF03061	PF03835	PF04716	PF05458	PF06470	PF07749	PF08540
PF00153	PF00485	PF00883	PF01267	PF01834	PF02390	PF03062	PF03836	PF04718	PF05460	PF06472	PF07763	PF08542
PF00155	PF00487	PF00884	PF01268	PF01835	PF02391	PF03068	PF03847	PF04719	PF05461	PF06479	PF07766	PF08544
PF00156	PF00488	PF00886	PF01269	PF01839	PF02394	PF03073	PF03849	PF04724	PF05463	PF06480	PF07767	
PF00157	PF00489	PF00887	PF01271	PF01840	PF02403	PF03074	PF03850	PF04727	PF05466	PF06482	PF07776	
PF00159	PF00490	PF00889	PF01273	PF01841	PF02404	PF03081	PF03868	PF04731	PF05470	PF06487	PF07778	
PF00160	PF00491	PF00890	PF01275	PF01842	PF02412	PF03088	PF03870	PF04732	PF05474	PF06512	PF07782	
PF00162	PF00493	PF00891	PF01279	PF01843	PF02423	PF03089	PF03871	PF04733	PF05477	PF06524	PF07807	
PF00163	PF00494	PF00895	PF01280	PF01846	PF02434	PF03091	PF03874	PF04739	PF05478	PF06534	PF07808	
PF00166	PF00498	PF00896	PF01281	PF01847	PF02436	PF03095	PF03875	PF04749	PF05483	PF06535	PF07810	
PF00169	PF00499	PF00899	PF01282	PF01852	PF02437	PF03096	PF03876	PF04750	PF05485	PF06540	PF07815	
PF00171	PF00503	PF00900	PF01283	PF01853	PF02441	PF03097	PF03896	PF04756	PF05486	PF06544	PF07830	
PF00173	PF00507	PF00903	PF01284	PF01857	PF02450	PF03099	PF03900	PF04757	PF05493	PF06546	PF07831	
PF00174	PF00510	PF00904	PF01285	PF01858	PF02453	PF03102	PF03901	PF04758	PF05495	PF06553	PF07834	
PF00175	PF00520	PF00907	PF01286	PF01868	PF02460	PF03104	PF03908	PF04762	PF05499	PF06554	PF07837	
PF00177	PF00521	PF00909	PF01287	PF01869	PF02466	PF03109	PF03909	PF04768	PF05502	PF06566	PF07842	
PF00178	PF00525	PF00916	PF01290	PF01871	PF02469	PF03114	PF03911	PF04774	PF05507	PF06567	PF07847	
PF00179	PF00530	PF00917	PF01291	PF01873	PF02485	PF03127	PF03914	PF04775	PF05510	PF06573	PF07850	
PF00180	PF00531	PF00918	PF01294	PF01875	PF02487	PF03128	PF03919	PF04777	PF05511	PF06581	PF07859	
PF00181	PF00534	PF00919	PF01299	PF01876	PF02493	PF03129	PF03920	PF04790	PF05517	PF06583	PF07883	
PF00184	PF00538	PF00923	PF01302	PF01896	PF02494	PF03131	PF03921	PF04795	PF05529	PF06588	PF07884	
PF00185	PF00542	PF00928	PF01315	PF01900	PF02513	PF03133	PF03931	PF04799	PF05557	PF06602	PF07885	
PF00186	PF00549	PF00929	PF01323	PF01907	PF02515	PF03134	PF03932	PF04800	PF05571	PF06607	PF07894	
PF00188	PF00550	PF00930	PF01327	PF01909	PF02516	PF03143	PF03939	PF04801	PF05572	PF06608	PF07904	
PF00189	PF00551	PF00931	PF01329	PF01912	PF02517	PF03144	PF03941	PF04803	PF05577	PF06614	PF07912	
PF00191	PF00552	PF00935	PF01331	PF01916	PF02525	PF03145	PF03943	PF04810	PF05586	PF06617	PF07915	
PF00193	PF00554	PF00939	PF01335	PF01918	PF02535	PF03146	PF03946	PF04811	PF05587	PF06621	PF07923	
PF00194	PF00557	PF00940	PF01336	PF01922	PF02536	PF03148	PF03947	PF04812	PF05600	PF06623	PF07926	
PF00198	PF00560	PF00941	PF01342	PF01923	PF02538	PF03149	PF03949	PF04813	PF05604	PF06625	PF07928	
PF00199	PF00562	PF00953	PF01347	PF01926	PF02544	PF03151	PF03950	PF04814	PF05622	PF06631	PF07933	

PF00201	PF00564	PF00955	PF01351	PF01928	PF02545	PF03152	PF03952	PF04815	PF05640	PF06632	PF07934
PF00202	PF00566	PF00956	PF01354	PF01929	PF02567	PF03153	PF03953	PF04818	PF05641	PF06637	PF07941
PF00203	PF00568	PF00957	PF01363	PF01937	PF02574	PF03155	PF03954	PF04819	PF05645	PF06638	PF07942
PF00204	PF00569	PF00958	PF01365	PF01938	PF02578	PF03159	PF03957	PF04821	PF05648	PF06645	PF07959
PF00205	PF00570	PF00962	PF01369	PF01963	PF02580	PF03160	PF03966	PF04822	PF05649	PF06650	PF07965
PF00206	PF00571	PF00965	PF01379	PF01965	PF02585	PF03165	PF03981	PF04824	PF05652	PF06657	PF07966
PF00207	PF00572	PF00969	PF01380	PF01966	PF02594	PF03166	PF03991	PF04825	PF05653	PF06662	PF07967
PF00208	PF00573	PF00970	PF01381	PF01967	PF02597	PF03167	PF03992	PF04831	PF05669	PF06663	PF07970
PF00209	PF00574	PF00975	PF01387	PF01968	PF02602	PF03171	PF03997	PF04836	PF05670	PF06668	PF07973
PF00210	PF00575	PF00976	PF01388	PF01974	PF02607	PF03172	PF03998	PF04840	PF05676	PF06677	PF07974
PF00211	PF00576	PF00984	PF01390	PF01979	PF02617	PF03177	PF04003	PF04841	PF05679	PF06699	PF07975
PF00212	PF00578	PF00988	PF01392	PF01981	PF02628	PF03178	PF04006	PF04845	PF05680	PF06701	PF07978
PF00213	PF00579	PF00992	PF01393	PF01984	PF02629	PF03179	PF04031	PF04847	PF05682	PF06702	PF07985
PF00214	PF00581	PF00993	PF01394	PF01990	PF02630	PF03184	PF04032	PF04855	PF05693	PF06703	PF07986
PF00215	PF00584	PF00994	PF01396	PF01991	PF02637	PF03185	PF04037	PF04856	PF05694	PF06726	PF07988
PF00217	PF00586	PF00996	PF01398	PF01992	PF02666	PF03188	PF04042	PF04858	PF05699	PF06728	PF07989
PF00219	PF00587	PF00997	PF01399	PF01997	PF02668	PF03194	PF04044	PF04868	PF05700	PF06729	PF08005
PF00220	PF00590	PF00999	PF01400	PF02002	PF02671	PF03199	PF04045	PF04869	PF05706	PF06732	PF08016
PF00221	PF00591	PF01000	PF01401	PF02005	PF02678	PF03200	PF04046	PF04871	PF05712	PF06733	PF08022
PF00224	PF00594	PF01007	PF01403	PF02008	PF02686	PF03208	PF04047	PF04874	PF05715	PF06741	PF08030
PF00225	PF00596	PF01008	PF01404	PF02012	PF02709	PF03215	PF04049	PF04882	PF05719	PF06743	PF08033
PF00227	PF00605	PF01012	PF01408	PF02014	PF02724	PF03223	PF04050	PF04886	PF05721	PF06747	PF08035
PF00229	PF00607	PF01015	PF01409	PF02017	PF02727	PF03224	PF04051	PF04900	PF05724	PF06752	PF08038
PF00230	PF00609	PF01016	PF01410	PF02019	PF02728	PF03226	PF04053	PF04901	PF05726	PF06773	PF08039
PF00231	PF00610	PF01017	PF01411	PF02020	PF02729	PF03227	PF04055	PF04902	PF05729	PF06775	PF08040
PF00232	PF00611	PF01018	PF01412	PF02022	PF02731	PF03232	PF04056	PF04904	PF05731	PF06777	PF08059
PF00233	PF00612	PF01020	PF01413	PF02025	PF02732	PF03234	PF04057	PF04905	PF05735	PF06783	PF08060
PF00235	PF00613	PF01023	PF01414	PF02026	PF02735	PF03247	PF04062	PF04908	PF05739	PF06784	PF08061
PF00236	PF00614	PF01025	PF01415	PF02036	PF02736	PF03248	PF04065	PF04909	PF05741	PF06807	PF08062
PF00237	PF00616	PF01026	PF01416	PF02037	PF02737	PF03250	PF04068	PF04910	PF05743	PF06809	PF08063
PF00238	PF00617	PF01027	PF01417	PF02038	PF02738	PF03253	PF04072	PF04912	PF05746	PF06814	PF08064
PF00241	PF00618	PF01028	PF01419	PF02039	PF02740	PF03256	PF04081	PF04921	PF05753	PF06816	PF08065

PF00243	PF00619	PF01030	PF01421	PF02044	PF02743	PF03259	PF04083	PF04923	PF05756	PF06817	PF08066
PF00244	PF00620	PF01031	PF01422	PF02046	PF02744	PF03261	PF04084	PF04926	PF05760	PF06818	PF08067
PF00245	PF00621	PF01033	PF01424	PF02055	PF02747	PF03265	PF04086	PF04928	PF05761	PF06821	PF08068
PF00246	PF00622	PF01035	PF01425	PF02057	PF02749	PF03266	PF04088	PF04934	PF05764	PF06825	PF08069
PF00248	PF00623	PF01039	PF01426	PF02058	PF02750	PF03271	PF04089	PF04935	PF05773	PF06828	PF08070
PF00249	PF00627	PF01040	PF01428	PF02059	PF02751	PF03285	PF04091	PF04938	PF05778	PF06831	PF08071
PF00250	PF00629	PF01042	PF01429	PF02064	PF02752	PF03297	PF04095	PF04939	PF05781	PF06839	PF08072
PF00253	PF00631	PF01048	PF01431	PF02070	PF02755	PF03298	PF04096	PF04950	PF05782	PF06858	PF08073
PF00255	PF00632	PF01053	PF01432	PF02071	PF02757	PF03299	PF04097	PF04952	PF05783	PF06870	PF08074
PF00256	PF00633	PF01055	PF01434	PF02072	PF02758	PF03301	PF04098	PF04959	PF05786	PF06875	PF08075
PF00258	PF00634	PF01056	PF01435	PF02077	PF02759	PF03308	PF04099	PF04960	PF05790	PF06881	PF08079
PF00260	PF00635	PF01058	PF01436	PF02078	PF02760	PF03311	PF04103	PF04961	PF05793	PF06883	PF08080
PF00261	PF00636	PF01059	PF01437	PF02079	PF02761	PF03321	PF04104	PF04968	PF05804	PF06886	PF08081
PF00262	PF00637	PF01061	PF01448	PF02083	PF02762	PF03332	PF04111	PF04969	PF05805	PF06905	PF08082
PF00264	PF00638	PF01062	PF01451	PF02089	PF02765	PF03345	PF04113	PF04970	PF05806	PF06907	PF08083
PF00265	PF00639	PF01063	PF01459	PF02093	PF02769	PF03351	PF04114	PF04979	PF05817	PF06910	PF08084
PF00266	PF00642	PF01064	PF01462	PF02096	PF02770	PF03357	PF04117	PF04983	PF05821	PF06916	PF08122
PF00268	PF00643	PF01066	PF01463	PF02099	PF02771	PF03359	PF04118	PF04988	PF05822	PF06920	PF08123
PF00273	PF00644	PF01067	PF01464	PF02100	PF02775	PF03360	PF04121	PF04990	PF05824	PF06936	PF08127
PF00274	PF00645	PF01068	PF01465	PF02101	PF02776	PF03366	PF04124	PF04992	PF05825	PF06937	PF08142
PF00276	PF00646	PF01070	PF01466	PF02104	PF02777	PF03367	PF04127	PF04997	PF05826	PF06941	PF08143
PF00277	PF00647	PF01071	PF01467	PF02106	PF02779	PF03368	PF04128	PF05000	PF05827	PF06951	PF08144
PF00278	PF00648	PF01073	PF01469	PF02109	PF02780	PF03370	PF04130	PF05001	PF05832	PF06954	PF08145
PF00281	PF00653	PF01074	PF01470	PF02114	PF02781	PF03372	PF04132	PF05002	PF05835	PF06957	PF08146
PF00285	PF00654	PF01079	PF01472	PF02115	PF02782	PF03378	PF04133	PF05004	PF05843	PF06959	PF08147
PF00287	PF00656	PF01080	PF01477	PF02121	PF02784	PF03388	PF04135	PF05005	PF05856	PF06963	PF08148
PF00288	PF00657	PF01082	PF01479	PF02126	PF02785	PF03398	PF04136	PF05008	PF05863	PF06978	PF08149
PF00289	PF00658	PF01084	PF01480	PF02127	PF02786	PF03399	PF04137	PF05010	PF05871	PF06984	PF08150
PF00291	PF00659	PF01085	PF01483	PF02130	PF02787	PF03402	PF04140	PF05014	PF05873	PF06990	PF08151
PF00292	PF00662	PF01086	PF01485	PF02134	PF02789	PF03403	PF04142	PF05019	PF05875	PF06991	PF08152
PF00293	PF00665	PF01087	PF01490	PF02135	PF02790	PF03404	PF04144	PF05020	PF05881	PF07001	PF08153
PF00294	PF00666	PF01088	PF01491	PF02136	PF02791	PF03414	PF04145	PF05021	PF05889	PF07002	PF08154

PF00297	PF00667	PF01090	PF01496	PF02137	PF02792	PF03446	PF04146	PF05022	PF05890	PF07010	PF08155
PF00298	PF00670	PF01091	PF01501	PF02138	PF02793	PF03450	PF04147	PF05024	PF05902	PF07019	PF08156
PF00300	PF00673	PF01092	PF01504	PF02140	PF02798	PF03451	PF04152	PF05026	PF05903	PF07034	PF08157
PF00303	PF00675	PF01093	PF01505	PF02141	PF02799	PF03453	PF04153	PF05028	PF05918	PF07035	PF08159
PF00305	PF00676	PF01096	PF01509	PF02142	PF02800	PF03454	PF04157	PF05029	PF05922	PF07047	PF08160
PF00306	PF00679	PF01099	PF01510	PF02145	PF02801	PF03455	PF04177	PF05030	PF05923	PF07074	PF08161
PF00307	PF00682	PF01101	PF01512	PF02146	PF02803	PF03456	PF04191	PF05033	PF05924	PF07084	PF08162
PF00310	PF00683	PF01103	PF01513	PF02148	PF02806	PF03462	PF04192	PF05038	PF05937	PF07093	PF08163
PF00312	PF00684	PF01105	PF01529	PF02149	PF02807	PF03463	PF04193	PF05039	PF05956	PF07111	PF08164
PF00313	PF00685	PF01106	PF01530	PF02150	PF02809	PF03464	PF04194	PF05040	PF05964	PF07140	PF08165
PF00316	PF00686	PF01109	PF01531	PF02157	PF02812	PF03465	PF04201	PF05041	PF05965	PF07142	PF08168
PF00317	PF00687	PF01110	PF01532	PF02159	PF02815	PF03467	PF04209	PF05044	PF05967	PF07145	PF08169
PF00318	PF00688	PF01111	PF01534	PF02161	PF02816	PF03477	PF04212	PF05046	PF05972	PF07147	PF08170
PF00319	PF00689	PF01112	PF01535	PF02165	PF02817	PF03483	PF04218	PF05047	PF05978	PF07156	PF08180
PF00320	PF00690	PF01114	PF01536	PF02167	PF02820	PF03484	PF04253	PF05051	PF05983	PF07159	PF08185
PF00322	PF00692	PF01115	PF01545	PF02170	PF02822	PF03485	PF04258	PF05053	PF05995	PF07163	PF08198

Vertebrate Pfam ID												
PF00005	PF00342	PF00753	PF01162	PF01633	PF02229	PF02910	PF03630	PF04548	PF05240	PF06248	PF07524	PF08327
PF00006	PF00343	PF00755	PF01163	PF01642	PF02230	PF02911	PF03635	PF04549	PF05241	PF06268	PF07525	PF08332
PF00009	PF00344	PF00756	PF01166	PF01652	PF02233	PF02913	PF03637	PF04553	PF05242	PF06292	PF07527	PF08333
PF00010	PF00346	PF00758	PF01167	PF01655	PF02234	PF02919	PF03643	PF04557	PF05250	PF06294	PF07528	PF08336
PF00013	PF00347	PF00762	PF01168	PF01661	PF02237	PF02920	PF03645	PF04558	PF05251	PF06297	PF07529	PF08337
PF00014	PF00348	PF00763	PF01169	PF01663	PF02238	PF02921	PF03647	PF04560	PF05253	PF06309	PF07531	PF08344
PF00017	PF00349	PF00764	PF01170	PF01667	PF02244	PF02922	PF03651	PF04561	PF05254	PF06311	PF07533	PF08351
PF00019	PF00350	PF00766	PF01172	PF01669	PF02245	PF02928	PF03660	PF04565	PF05255	PF06312	PF07539	PF08357
PF00024	PF00351	PF00769	PF01175	PF01687	PF02251	PF02931	PF03661	PF04566	PF05276	PF06320	PF07540	PF08365
PF00025	PF00352	PF00773	PF01176	PF01694	PF02252	PF02932	PF03665	PF04567	PF05279	PF06327	PF07541	PF08368
PF00026	PF00354	PF00774	PF01179	PF01699	PF02257	PF02933	PF03666	PF04568	PF05281	PF06328	PF07542	PF08373
PF00027	PF00355	PF00777	PF01180	PF01702	PF02259	PF02934	PF03669	PF04571	PF05282	PF06331	PF07545	PF08374
PF00029	PF00357	PF00778	PF01182	PF01704	PF02260	PF02935	PF03670	PF04572	PF05283	PF06333	PF07546	PF08375
PF00030	PF00362	PF00779	PF01187	PF01709	PF02262	PF02936	PF03671	PF04573	PF05287	PF06337	PF07562	PF08377
PF00032	PF00364	PF00780	PF01189	PF01712	PF02267	PF02937	PF03676	PF04579	PF05291	PF06345	PF07565	PF08383
PF00034	PF00365	PF00781	PF01191	PF01713	PF02268	PF02939	PF03690	PF04587	PF05292	PF06350	PF07569	PF08384
PF00037	PF00366	PF00782	PF01192	PF01715	PF02269	PF02946	PF03694	PF04588	PF05296	PF06365	PF07571	PF08385
PF00039	PF00368	PF00784	PF01194	PF01722	PF02270	PF02947	PF03700	PF04589	PF05298	PF06367	PF07576	PF08389
PF00040	PF00370	PF00786	PF01198	PF01725	PF02271	PF02948	PF03712	PF04592	PF05300	PF06372	PF07593	PF08390
PF00044	PF00375	PF00787	PF01199	PF01729	PF02272	PF02953	PF03715	PF04593	PF05303	PF06373	PF07646	PF08393
PF00045	PF00377	PF00789	PF01200	PF01731	PF02274	PF02961	PF03719	PF04597	PF05324	PF06374	PF07647	PF08397
PF00046	PF00378	PF00790	PF01201	PF01733	PF02275	PF02965	PF03720	PF04598	PF05326	PF06375	PF07649	PF08399
PF00048	PF00382	PF00791	PF01202	PF01734	PF02284	PF02966	PF03721	PF04614	PF05328	PF06384	PF07650	PF08403
PF00050	PF00383	PF00792	PF01204	PF01735	PF02285	PF02969	PF03725	PF04615	PF05337	PF06387	PF07651	PF08412
PF00051	PF00384	PF00795	PF01207	PF01740	PF02290	PF02970	PF03726	PF04617	PF05347	PF06391	PF07657	PF08416
PF00052	PF00385	PF00797	PF01208	PF01743	PF02291	PF02971	PF03727	PF04621	PF05349	PF06393	PF07662	PF08418
PF00054	PF00386	PF00801	PF01210	PF01747	PF02296	PF02978	PF03730	PF04627	PF05350	PF06396	PF07670	PF08420
PF00055	PF00388	PF00804	PF01213	PF01749	PF02301	PF02984	PF03731	PF04628	PF05351	PF06399	PF07677	PF08423
PF00056	PF00390	PF00806	PF01214	PF01751	PF02310	PF02985	PF03735	PF04629	PF05355	PF06400	PF07678	PF08424

PF00058	PF00393	PF00808	PF01215	PF01753	PF02312	PF02990	PF03736	PF04636	PF05362	PF06401	PF07684	PF08429
PF00060	PF00394	PF00809	PF01216	PF01754	PF02318	PF02991	PF03747	PF04641	PF05365	PF06403	PF07687	PF08430
PF00062	PF00396	PF00810	PF01217	PF01756	PF02319	PF02994	PF03758	PF04643	PF05366	PF06418	PF07699	PF08433
PF00066	PF00397	PF00811	PF01218	PF01758	PF02320	PF02996	PF03762	PF04652	PF05378	PF06419	PF07701	PF08434
PF00068	PF00398	PF00812	PF01221	PF01759	PF02330	PF03002	PF03764	PF04658	PF05383	PF06424	PF07703	PF08441
PF00072	PF00401	PF00814	PF01222	PF01762	PF02337	PF03006	PF03765	PF04664	PF05386	PF06427	PF07706	PF08442
PF00074	PF00402	PF00817	PF01223	PF01765	PF02338	PF03009	PF03770	PF04667	PF05392	PF06428	PF07710	PF08447
PF00075	PF00403	PF00819	PF01227	PF01770	PF02344	PF03016	PF03781	PF04675	PF05395	PF06432	PF07711	PF08449
PF00077	PF00408	PF00821	PF01229	PF01773	PF02345	PF03020	PF03792	PF04679	PF05399	PF06441	PF07716	PF08450
PF00080	PF00410	PF00822	PF01230	PF01775	PF02347	PF03024	PF03798	PF04680	PF05404	PF06446	PF07717	PF08451
PF00081	PF00411	PF00826	PF01231	PF01776	PF02348	PF03028	PF03803	PF04683	PF05405	PF06448	PF07718	PF08454
PF00083	PF00412	PF00827	PF01233	PF01777	PF02350	PF03029	PF03807	PF04691	PF05406	PF06455	PF07719	PF08466
PF00084	PF00414	PF00828	PF01234	PF01778	PF02351	PF03031	PF03810	PF04692	PF05418	PF06456	PF07720	PF08473
PF00085	PF00416	PF00831	PF01237	PF01779	PF02354	PF03032	PF03813	PF04695	PF05422	PF06459	PF07722	PF08474
PF00087	PF00418	PF00832	PF01238	PF01780	PF02359	PF03034	PF03815	PF04696	PF05427	PF06461	PF07723	PF08477
PF00088	PF00420	PF00833	PF01239	PF01781	PF02366	PF03036	PF03820	PF04697	PF05428	PF06462	PF07724	PF08487
PF00090	PF00428	PF00834	PF01242	PF01783	PF02372	PF03039	PF03821	PF04698	PF05434	PF06463	PF07731	PF08490
PF00091	PF00433	PF00835	PF01243	PF01784	PF02373	PF03045	PF03823	PF04699	PF05438	PF06464	PF07738	PF08491
PF00092	PF00436	PF00836	PF01244	PF01791	PF02374	PF03051	PF03826	PF04704	PF05439	PF06465	PF07741	PF08492
PF00093	PF00439	PF00837	PF01245	PF01794	PF02375	PF03054	PF03827	PF04706	PF05450	PF06466	PF07742	PF08499
PF00094	PF00441	PF00838	PF01246	PF01798	PF02376	PF03055	PF03828	PF04707	PF05454	PF06467	PF07743	PF08504
PF00095	PF00444	PF00839	PF01247	PF01799	PF02377	PF03056	PF03832	PF04709	PF05456	PF06468	PF07744	PF08506
PF00096	PF00445	PF00850	PF01248	PF01803	PF02383	PF03061	PF03834	PF04710	PF05458	PF06469	PF07748	PF08510
PF00098	PF00446	PF00852	PF01249	PF01805	PF02389	PF03062	PF03835	PF04711	PF05460	PF06470	PF07749	PF08513
PF00105	PF00447	PF00853	PF01251	PF01808	PF02390	PF03068	PF03836	PF04716	PF05461	PF06472	PF07763	PF08514
PF00106	PF00448	PF00855	PF01253	PF01812	PF02391	PF03073	PF03847	PF04718	PF05463	PF06479	PF07766	PF08515
PF00107	PF00450	PF00856	PF01254	PF01813	PF02403	PF03074	PF03849	PF04719	PF05466	PF06480	PF07767	PF08516
PF00108	PF00452	PF00858	PF01255	PF01821	PF02404	PF03081	PF03850	PF04724	PF05470	PF06482	PF07776	PF08517
PF00109	PF00454	PF00860	PF01257	PF01822	PF02412	PF03088	PF03868	PF04727	PF05477	PF06487	PF07778	PF08518
PF00111	PF00458	PF00861	PF01259	PF01823	PF02422	PF03089	PF03870	PF04731	PF05478	PF06512	PF07782	PF08519
PF00112	PF00459	PF00864	PF01262	PF01833	PF02423	PF03091	PF03871	PF04732	PF05483	PF06524	PF07807	PF08523
PF00113	PF00462	PF00865	PF01263	PF01835	PF02434	PF03095	PF03874	PF04733	PF05485	PF06534	PF07808	PF08526

PF00115	PF00464	PF00867	PF01265	PF01839	PF02436	PF03096	PF03875	PF04739	PF05486	PF06535	PF07810	PF08527
PF00116	PF00466	PF00868	PF01267	PF01840	PF02437	PF03097	PF03876	PF04749	PF05493	PF06540	PF07815	PF08534
PF00117	PF00467	PF00870	PF01268	PF01841	PF02441	PF03099	PF03896	PF04750	PF05495	PF06544	PF07818	PF08540
PF00118	PF00468	PF00875	PF01269	PF01842	PF02450	PF03102	PF03900	PF04756	PF05499	PF06546	PF07830	PF08542
PF00119	PF00472	PF00879	PF01271	PF01843	PF02453	PF03104	PF03901	PF04757	PF05502	PF06553	PF07831	PF08544
PF00120	PF00473	PF00883	PF01273	PF01846	PF02460	PF03109	PF03908	PF04758	PF05507	PF06554	PF07834	
PF00121	PF00474	PF00884	PF01275	PF01847	PF02466	PF03127	PF03909	PF04762	PF05510	PF06566	PF07837	
PF00123	PF00476	PF00886	PF01279	PF01852	PF02469	PF03128	PF03911	PF04768	PF05517	PF06567	PF07842	
PF00125	PF00479	PF00887	PF01280	PF01853	PF02485	PF03129	PF03914	PF04774	PF05529	PF06573	PF07847	
PF00129	PF00480	PF00889	PF01281	PF01857	PF02487	PF03131	PF03919	PF04775	PF05557	PF06581	PF07850	
PF00130	PF00481	PF00890	PF01282	PF01858	PF02493	PF03133	PF03920	PF04777	PF05571	PF06583	PF07859	
PF00131	PF00485	PF00891	PF01283	PF01868	PF02494	PF03134	PF03921	PF04790	PF05572	PF06588	PF07883	
PF00132	PF00487	PF00895	PF01284	PF01869	PF02513	PF03143	PF03931	PF04795	PF05577	PF06602	PF07884	
PF00135	PF00488	PF00896	PF01285	PF01871	PF02515	PF03144	PF03932	PF04799	PF05586	PF06608	PF07885	
PF00136	PF00489	PF00899	PF01286	PF01873	PF02516	PF03145	PF03939	PF04800	PF05587	PF06614	PF07894	
PF00137	PF00490	PF00900	PF01287	PF01875	PF02517	PF03146	PF03940	PF04801	PF05600	PF06621	PF07904	
PF00144	PF00491	PF00903	PF01290	PF01876	PF02525	PF03148	PF03941	PF04803	PF05604	PF06623	PF07912	
PF00145	PF00493	PF00904	PF01291	PF01896	PF02535	PF03149	PF03943	PF04810	PF05609	PF06625	PF07915	
PF00146	PF00494	PF00907	PF01294	PF01900	PF02536	PF03151	PF03946	PF04811	PF05622	PF06631	PF07923	
PF00149	PF00498	PF00909	PF01299	PF01907	PF02538	PF03152	PF03947	PF04812	PF05640	PF06632	PF07926	
PF00151	PF00499	PF00916	PF01302	PF01909	PF02544	PF03153	PF03949	PF04813	PF05641	PF06637	PF07928	
PF00152	PF00503	PF00917	PF01315	PF01912	PF02545	PF03155	PF03950	PF04814	PF05645	PF06638	PF07933	
PF00153	PF00507	PF00919	PF01323	PF01916	PF02567	PF03159	PF03952	PF04815	PF05648	PF06645	PF07934	
PF00156	PF00510	PF00923	PF01327	PF01918	PF02574	PF03160	PF03953	PF04818	PF05649	PF06650	PF07941	
PF00157	PF00521	PF00928	PF01329	PF01922	PF02578	PF03165	PF03954	PF04819	PF05652	PF06657	PF07942	
PF00159	PF00525	PF00929	PF01331	PF01923	PF02580	PF03166	PF03957	PF04821	PF05653	PF06662	PF07959	
PF00160	PF00531	PF00930	PF01335	PF01926	PF02585	PF03167	PF03966	PF04822	PF05669	PF06663	PF07965	
PF00162	PF00534	PF00931	PF01336	PF01928	PF02594	PF03171	PF03981	PF04824	PF05670	PF06668	PF07966	
PF00163	PF00538	PF00932	PF01342	PF01929	PF02597	PF03172	PF03991	PF04825	PF05676	PF06677	PF07967	
PF00166	PF00542	PF00935	PF01347	PF01937	PF02602	PF03177	PF03992	PF04831	PF05679	PF06699	PF07970	
PF00169	PF00549	PF00939	PF01351	PF01938	PF02607	PF03178	PF03997	PF04836	PF05680	PF06701	PF07973	
PF00171	PF00550	PF00940	PF01354	PF01958	PF02617	PF03179	PF03998	PF04840	PF05682	PF06702	PF07974	

PF00173	PF00551	PF00941	PF01363	PF01963	PF02628	PF03184	PF04003	PF04841	PF05693	PF06703	PF07975
PF00174	PF00552	PF00953	PF01365	PF01965	PF02629	PF03185	PF04006	PF04845	PF05694	PF06726	PF07978
PF00175	PF00554	PF00955	PF01369	PF01966	PF02630	PF03188	PF04030	PF04847	PF05699	PF06728	PF07985
PF00177	PF00557	PF00956	PF01379	PF01967	PF02637	PF03194	PF04031	PF04855	PF05700	PF06729	PF07986
PF00178	PF00560	PF00957	PF01380	PF01968	PF02666	PF03199	PF04032	PF04856	PF05706	PF06732	PF07988
PF00179	PF00562	PF00958	PF01381	PF01974	PF02668	PF03200	PF04037	PF04858	PF05712	PF06733	PF07989
PF00180	PF00564	PF00962	PF01382	PF01979	PF02671	PF03208	PF04042	PF04868	PF05715	PF06741	PF08005
PF00181	PF00566	PF00965	PF01387	PF01981	PF02678	PF03215	PF04044	PF04869	PF05719	PF06743	PF08016
PF00184	PF00568	PF00969	PF01388	PF01984	PF02709	PF03223	PF04045	PF04871	PF05721	PF06747	PF08022
PF00185	PF00569	PF00970	PF01390	PF01990	PF02724	PF03224	PF04046	PF04874	PF05724	PF06752	PF08023
PF00186	PF00570	PF00975	PF01392	PF01991	PF02727	PF03226	PF04047	PF04882	PF05726	PF06753	PF08030
PF00188	PF00571	PF00984	PF01393	PF01992	PF02728	PF03227	PF04049	PF04886	PF05729	PF06773	PF08033
PF00189	PF00572	PF00988	PF01394	PF01997	PF02729	PF03232	PF04050	PF04900	PF05731	PF06775	PF08035
PF00191	PF00573	PF00992	PF01396	PF02002	PF02731	PF03234	PF04051	PF04901	PF05735	PF06777	PF08038
PF00193	PF00574	PF00993	PF01398	PF02005	PF02732	PF03247	PF04053	PF04902	PF05739	PF06783	PF08039
PF00194	PF00575	PF00994	PF01399	PF02008	PF02735	PF03248	PF04055	PF04904	PF05741	PF06784	PF08040
PF00198	PF00576	PF00996	PF01400	PF02012	PF02736	PF03250	PF04056	PF04905	PF05742	PF06807	PF08059
PF00199	PF00578	PF00999	PF01401	PF02014	PF02737	PF03253	PF04057	PF04908	PF05743	PF06809	PF08060
PF00201	PF00579	PF01000	PF01403	PF02017	PF02738	PF03256	PF04062	PF04909	PF05746	PF06814	PF08061
PF00202	PF00581	PF01007	PF01404	PF02019	PF02740	PF03259	PF04065	PF04910	PF05753	PF06816	PF08062
PF00203	PF00583	PF01008	PF01408	PF02020	PF02743	PF03261	PF04068	PF04912	PF05756	PF06817	PF08063
PF00204	PF00584	PF01012	PF01409	PF02022	PF02744	PF03265	PF04072	PF04921	PF05760	PF06818	PF08064
PF00205	PF00586	PF01014	PF01410	PF02025	PF02747	PF03266	PF04081	PF04923	PF05761	PF06821	PF08065
PF00206	PF00587	PF01015	PF01411	PF02026	PF02749	PF03271	PF04083	PF04926	PF05764	PF06825	PF08066
PF00207	PF00590	PF01016	PF01412	PF02036	PF02750	PF03285	PF04084	PF04934	PF05773	PF06828	PF08067
PF00208	PF00591	PF01017	PF01413	PF02037	PF02751	PF03297	PF04086	PF04935	PF05778	PF06831	PF08068
PF00209	PF00594	PF01018	PF01414	PF02039	PF02752	PF03299	PF04088	PF04938	PF05781	PF06839	PF08069
PF00210	PF00596	PF01020	PF01415	PF02044	PF02755	PF03301	PF04089	PF04939	PF05782	PF06858	PF08070
PF00211	PF00605	PF01023	PF01416	PF02046	PF02757	PF03308	PF04091	PF04950	PF05783	PF06870	PF08071
PF00213	PF00609	PF01025	PF01417	PF02055	PF02758	PF03311	PF04095	PF04952	PF05786	PF06875	PF08072
PF00214	PF00610	PF01026	PF01419	PF02057	PF02759	PF03321	PF04096	PF04959	PF05790	PF06881	PF08073
PF00215	PF00611	PF01027	PF01421	PF02058	PF02760	PF03332	PF04097	PF04960	PF05793	PF06886	PF08074

PF00217	PF00612	PF01028	PF01422	PF02059	PF02761	PF03345	PF04098	PF04961	PF05794	PF06905	PF08075
PF00219	PF00613	PF01030	PF01424	PF02064	PF02762	PF03351	PF04099	PF04968	PF05804	PF06907	PF08079
PF00220	PF00614	PF01031	PF01425	PF02071	PF02765	PF03357	PF04103	PF04969	PF05805	PF06910	PF08080
PF00221	PF00616	PF01033	PF01426	PF02077	PF02769	PF03359	PF04104	PF04970	PF05806	PF06916	PF08081
PF00224	PF00617	PF01035	PF01428	PF02078	PF02770	PF03360	PF04111	PF04979	PF05808	PF06920	PF08082
PF00225	PF00618	PF01039	PF01431	PF02079	PF02771	PF03366	PF04113	PF04983	PF05817	PF06936	PF08083
PF00227	PF00620	PF01040	PF01432	PF02083	PF02775	PF03367	PF04114	PF04988	PF05821	PF06937	PF08084
PF00230	PF00621	PF01042	PF01434	PF02089	PF02776	PF03368	PF04117	PF04990	PF05822	PF06941	PF08103
PF00231	PF00623	PF01048	PF01435	PF02093	PF02777	PF03370	PF04118	PF04992	PF05824	PF06951	PF08110
PF00232	PF00626	PF01053	PF01436	PF02096	PF02779	PF03372	PF04121	PF05000	PF05826	PF06954	PF08121
PF00233	PF00627	PF01055	PF01437	PF02099	PF02780	PF03378	PF04124	PF05001	PF05827	PF06957	PF08122
PF00235	PF00629	PF01058	PF01448	PF02100	PF02781	PF03388	PF04127	PF05002	PF05832	PF06959	PF08123
PF00236	PF00631	PF01059	PF01451	PF02101	PF02782	PF03398	PF04128	PF05004	PF05835	PF06963	PF08127
PF00237	PF00632	PF01061	PF01459	PF02104	PF02784	PF03399	PF04130	PF05005	PF05843	PF06970	PF08131
PF00238	PF00635	PF01062	PF01462	PF02106	PF02785	PF03402	PF04132	PF05008	PF05856	PF06978	PF08142
PF00241	PF00636	PF01063	PF01463	PF02109	PF02786	PF03403	PF04133	PF05010	PF05863	PF06984	PF08143
PF00243	PF00637	PF01064	PF01464	PF02114	PF02787	PF03404	PF04135	PF05014	PF05871	PF06990	PF08144
PF00244	PF00638	PF01066	PF01465	PF02115	PF02789	PF03414	PF04136	PF05019	PF05873	PF06991	PF08145
PF00245	PF00639	PF01067	PF01466	PF02121	PF02791	PF03441	PF04137	PF05020	PF05875	PF06994	PF08146
PF00246	PF00642	PF01068	PF01467	PF02126	PF02792	PF03446	PF04140	PF05021	PF05881	PF07001	PF08147
PF00248	PF00643	PF01070	PF01469	PF02127	PF02793	PF03447	PF04142	PF05022	PF05889	PF07002	PF08148
PF00249	PF00644	PF01071	PF01470	PF02130	PF02799	PF03450	PF04144	PF05024	PF05890	PF07010	PF08149
PF00250	PF00645	PF01073	PF01472	PF02134	PF02800	PF03451	PF04145	PF05026	PF05902	PF07019	PF08150
PF00253	PF00646	PF01074	PF01477	PF02135	PF02801	PF03453	PF04146	PF05028	PF05903	PF07034	PF08151
PF00255	PF00647	PF01079	PF01479	PF02136	PF02803	PF03454	PF04147	PF05029	PF05918	PF07035	PF08152
PF00256	PF00648	PF01080	PF01480	PF02137	PF02806	PF03455	PF04152	PF05030	PF05922	PF07047	PF08153
PF00258	PF00653	PF01082	PF01483	PF02138	PF02807	PF03456	PF04153	PF05033	PF05923	PF07074	PF08154
PF00260	PF00654	PF01084	PF01485	PF02140	PF02809	PF03462	PF04157	PF05038	PF05924	PF07084	PF08155
PF00261	PF00656	PF01085	PF01490	PF02141	PF02812	PF03463	PF04177	PF05039	PF05937	PF07093	PF08156
PF00262	PF00657	PF01086	PF01491	PF02142	PF02815	PF03464	PF04191	PF05040	PF05956	PF07111	PF08157
PF00264	PF00658	PF01087	PF01496	PF02146	PF02816	PF03465	PF04192	PF05041	PF05964	PF07140	PF08159
PF00265	PF00659	PF01088	PF01501	PF02148	PF02817	PF03467	PF04193	PF05044	PF05965	PF07142	PF08160

PF00266	PF00662	PF01090	PF01504	PF02149	PF02820	PF03477	PF04194	PF05046	PF05967	PF07145	PF08161
PF00268	PF00666	PF01091	PF01505	PF02150	PF02822	PF03483	PF04201	PF05047	PF05972	PF07147	PF08162
PF00271	PF00667	PF01092	PF01509	PF02157	PF02823	PF03484	PF04209	PF05051	PF05978	PF07156	PF08163
PF00273	PF00670	PF01093	PF01510	PF02165	PF02824	PF03485	PF04212	PF05053	PF05983	PF07159	PF08164
PF00274	PF00673	PF01096	PF01512	PF02167	PF02825	PF03487	PF04218	PF05060	PF05995	PF07163	PF08165
PF00276	PF00675	PF01101	PF01513	PF02170	PF02826	PF03489	PF04253	PF05063	PF05997	PF07169	PF08168
PF00278	PF00676	PF01103	PF01529	PF02171	PF02827	PF03491	PF04258	PF05064	PF06001	PF07175	PF08169
PF00281	PF00679	PF01105	PF01530	PF02172	PF02828	PF03493	PF04263	PF05071	PF06008	PF07177	PF08170
PF00285	PF00682	PF01106	PF01531	PF02173	PF02836	PF03494	PF04265	PF05083	PF06009	PF07189	PF08180
PF00287	PF00683	PF01109	PF01532	PF02174	PF02837	PF03501	PF04272	PF05089	PF06010	PF07202	PF08185
PF00288	PF00684	PF01110	PF01534	PF02177	PF02838	PF03508	PF04275	PF05090	PF06012	PF07225	PF08188
PF00289	PF00685	PF01111	PF01536	PF02178	PF02840	PF03509	PF04280	PF05091	PF06017	PF07260	PF08189
PF00291	PF00686	PF01112	PF01545	PF02179	PF02841	PF03511	PF04300	PF05093	PF06025	PF07263	PF08198
PF00292	PF00687	PF01114	PF01549	PF02180	PF02843	PF03516	PF04376	PF05104	PF06026	PF07264	PF08202
PF00293	PF00689	PF01115	PF01551	PF02181	PF02844	PF03517	PF04377	PF05110	PF06031	PF07276	PF08204
PF00294	PF00690	PF01119	PF01553	PF02182	PF02845	PF03520	PF04379	PF05111	PF06046	PF07281	PF08210
PF00297	PF00692	PF01120	PF01556	PF02184	PF02847	PF03521	PF04382	PF05118	PF06049	PF07286	PF08212
PF00298	PF00694	PF01121	PF01557	PF02185	PF02852	PF03522	PF04388	PF05127	PF06052	PF07292	PF08213
PF00300	PF00696	PF01122	PF01562	PF02186	PF02854	PF03523	PF04389	PF05131	PF06058	PF07297	PF08215
PF00303	PF00698	PF01124	PF01564	PF02187	PF02862	PF03528	PF04406	PF05132	PF06060	PF07303	PF08216
PF00305	PF00704	PF01125	PF01565	PF02189	PF02864	PF03529	PF04408	PF05147	PF06068	PF07324	PF08221
PF00306	PF00705	PF01126	PF01566	PF02190	PF02865	PF03530	PF04410	PF05148	PF06071	PF07334	PF08225
PF00307	PF00707	PF01127	PF01569	PF02191	PF02866	PF03531	PF04418	PF05153	PF06079	PF07347	PF08232
PF00310	PF00708	PF01129	PF01571	PF02192	PF02867	PF03533	PF04419	PF05158	PF06083	PF07353	PF08235
PF00312	PF00709	PF01130	PF01575	PF02194	PF02870	PF03535	PF04420	PF05160	PF06087	PF07354	PF08240
PF00313	PF00710	PF01131	PF01582	PF02196	PF02872	PF03540	PF04421	PF05162	PF06093	PF07359	PF08241
PF00316	PF00711	PF01133	PF01583	PF02197	PF02874	PF03542	PF04423	PF05181	PF06105	PF07400	PF08242
PF00317	PF00714	PF01134	PF01585	PF02198	PF02875	PF03546	PF04427	PF05184	PF06119	PF07406	PF08246
PF00318	PF00715	PF01135	PF01586	PF02199	PF02877	PF03547	PF04433	PF05185	PF06121	PF07412	PF08248
PF00319	PF00717	PF01138	PF01588	PF02201	PF02879	PF03556	PF04434	PF05186	PF06140	PF07421	PF08256
PF00320	PF00719	PF01140	PF01590	PF02202	PF02880	PF03561	PF04437	PF05187	PF06148	PF07425	PF08264
PF00323	PF00727	PF01142	PF01591	PF02204	PF02881	PF03567	PF04438	PF05188	PF06179	PF07440	PF08265

PF00324	PF00728	PF01144	PF01592	PF02205	PF02882	PF03568	PF04440	PF05189	PF06189	PF07443	PF08266
PF00326	PF00731	PF01146	PF01596	PF02207	PF02883	PF03571	PF04442	PF05190	PF06201	PF07448	PF08271
PF00327	PF00732	PF01148	PF01597	PF02208	PF02885	PF03572	PF04487	PF05191	PF06202	PF07452	PF08279
PF00328	PF00733	PF01150	PF01598	PF02209	PF02887	PF03575	PF04488	PF05193	PF06209	PF07458	PF08285
PF00329	PF00735	PF01151	PF01599	PF02212	PF02888	PF03577	PF04494	PF05195	PF06211	PF07469	PF08288
PF00330	PF00736	PF01153	PF01602	PF02213	PF02889	PF03587	PF04495	PF05196	PF06212	PF07474	PF08292
PF00333	PF00743	PF01154	PF01603	PF02214	PF02891	PF03600	PF04503	PF05199	PF06214	PF07479	PF08295
PF00334	PF00748	PF01157	PF01608	PF02218	PF02892	PF03604	PF04505	PF05207	PF06218	PF07496	PF08311
PF00335	PF00749	PF01158	PF01619	PF02219	PF02893	PF03607	PF04511	PF05208	PF06220	PF07500	PF08313
PF00337	PF00750	PF01159	PF01624	PF02221	PF02897	PF03615	PF04515	PF05222	PF06221	PF07502	PF08318
PF00338	PF00751	PF01160	PF01625	PF02223	PF02898	PF03619	PF04538	PF05236	PF06229	PF07521	PF08321
PF00339	PF00752	PF01161	PF01630	PF02225	PF02902	PF03623	PF04547	PF05237	PF06237	PF07522	PF08326

Metazoa + Fungi Pfam ID														
PF00005	PF00398	PF00809	PF01187	PF01599	PF02140	PF02760	PF03260	PF03896	PF04500	PF05060	PF05781	PF06644	PF07711	PF08170
PF00006	PF00399	PF00810	PF01191	PF01602	PF02141	PF02761	PF03261	PF03900	PF04502	PF05063	PF05782	PF06645	PF07716	PF08171
PF00010	PF00401	PF00811	PF01192	PF01603	PF02144	PF02762	PF03265	PF03901	PF04503	PF05064	PF05783	PF06650	PF07717	PF08174
PF00013	PF00402	PF00812	PF01194	PF01608	PF02146	PF02765	PF03266	PF03902	PF04505	PF05071	PF05786	PF06652	PF07718	PF08180
PF00014	PF00403	PF00814	PF01195	PF01619	PF02148	PF02769	PF03271	PF03907	PF04506	PF05083	PF05790	PF06657	PF07719	PF08184
PF00017	PF00410	PF00815	PF01196	PF01625	PF02149	PF02770	PF03285	PF03908	PF04511	PF05089	PF05793	PF06658	PF07720	PF08185
PF00019	PF00411	PF00817	PF01199	PF01627	PF02150	PF02771	PF03291	PF03909	PF04515	PF05090	PF05794	PF06662	PF07722	PF08187
PF00025	PF00412	PF00819	PF01200	PF01630	PF02152	PF02772	PF03297	PF03911	PF04516	PF05091	PF05804	PF06663	PF07723	PF08188
PF00026	PF00414	PF00821	PF01201	PF01634	PF02156	PF02773	PF03299	PF03914	PF04538	PF05093	PF05805	PF06668	PF07724	PF08189
PF00029	PF00416	PF00822	PF01202	PF01642	PF02157	PF02774	PF03301	PF03915	PF04547	PF05104	PF05808	PF06677	PF07731	PF08192
PF00030	PF00418	PF00826	PF01204	PF01644	PF02162	PF02779	PF03308	PF03919	PF04548	PF05110	PF05811	PF06687	PF07732	PF08194
PF00032	PF00420	PF00827	PF01205	PF01645	PF02165	PF02780	PF03311	PF03921	PF04549	PF05111	PF05817	PF06699	PF07735	PF08198
PF00034	PF00425	PF00828	PF01207	PF01648	PF02167	PF02781	PF03321	PF03928	PF04553	PF05118	PF05821	PF06701	PF07738	PF08202
PF00037	PF00428	PF00829	PF01208	PF01652	PF02170	PF02784	PF03332	PF03931	PF04557	PF05127	PF05822	PF06702	PF07740	PF08203
PF00039	PF00433	PF00830	PF01209	PF01655	PF02171	PF02785	PF03344	PF03932	PF04558	PF05131	PF05823	PF06703	PF07741	PF08204
PF00040	PF00436	PF00831	PF01210	PF01658	PF02172	PF02789	PF03345	PF03935	PF04560	PF05132	PF05824	PF06726	PF07742	PF08206
PF00044	PF00438	PF00832	PF01212	PF01661	PF02173	PF02791	PF03348	PF03937	PF04561	PF05141	PF05826	PF06728	PF07743	PF08208
PF00045	PF00439	PF00833	PF01213	PF01663	PF02174	PF02792	PF03351	PF03939	PF04565	PF05147	PF05827	PF06729	PF07744	PF08210
PF00048	PF00444	PF00834	PF01214	PF01667	PF02177	PF02793	PF03359	PF03940	PF04566	PF05148	PF05832	PF06732	PF07745	PF08212
PF00050	PF00445	PF00835	PF01215	PF01669	PF02178	PF02795	PF03360	PF03941	PF04567	PF05151	PF05835	PF06733	PF07748	PF08213
PF00051	PF00446	PF00836	PF01216	PF01670	PF02179	PF02799	PF03366	PF03943	PF04568	PF05153	PF05836	PF06740	PF07749	PF08215
PF00052	PF00447	PF00837	PF01217	PF01679	PF02180	PF02800	PF03367	PF03946	PF04571	PF05154	PF05839	PF06741	PF07763	PF08216
PF00054	PF00448	PF00839	PF01218	PF01680	PF02181	PF02803	PF03368	PF03947	PF04572	PF05156	PF05841	PF06743	PF07766	PF08217
PF00055	PF00449	PF00840	PF01220	PF01681	PF02182	PF02804	PF03370	PF03949	PF04573	PF05158	PF05843	PF06752	PF07767	PF08221
PF00056	PF00450	PF00842	PF01221	PF01683	PF02184	PF02806	PF03371	PF03950	PF04574	PF05160	PF05849	PF06753	PF07770	PF08225
PF00058	PF00451	PF00849	PF01223	PF01687	PF02185	PF02807	PF03372	PF03952	PF04579	PF05161	PF05856	PF06756	PF07774	PF08226
PF00060	PF00452	PF00850	PF01226	PF01694	PF02186	PF02809	PF03378	PF03954	PF04587	PF05162	PF05859	PF06757	PF07776	PF08227
PF00062	PF00454	PF00852	PF01227	PF01697	PF02187	PF02811	PF03381	PF03957	PF04588	PF05172	PF05863	PF06766	PF07778	PF08228
PF00066	PF00457	PF00853	PF01229	PF01699	PF02189	PF02812	PF03388	PF03959	PF04589	PF05176	PF05871	PF06773	PF07780	PF08229

PF00068	PF00458	PF00855	PF01230	PF01702	PF02190	PF02815	PF03392	PF03964	PF04590	PF05178	PF05875	PF06775	PF07782	PF08232
PF00072	PF00459	PF00856	PF01231	PF01704	PF02191	PF02816	PF03398	PF03966	PF04592	PF05180	PF05879	PF06777	PF07798	PF08234
PF00074	PF00462	PF00857	PF01232	PF01705	PF02192	PF02817	PF03399	PF03969	PF04593	PF05181	PF05881	PF06783	PF07807	PF08235
PF00075	PF00463	PF00860	PF01233	PF01709	PF02194	PF02819	PF03402	PF03972	PF04597	PF05182	PF05889	PF06784	PF07808	PF08236
PF00077	PF00464	PF00861	PF01234	PF01712	PF02195	PF02820	PF03403	PF03973	PF04598	PF05183	PF05890	PF06807	PF07810	PF08240
PF00080	PF00465	PF00864	PF01237	PF01713	PF02196	PF02822	PF03404	PF03980	PF04603	PF05184	PF05891	PF06814	PF07815	PF08242
PF00081	PF00466	PF00865	PF01238	PF01715	PF02198	PF02823	PF03414	PF03981	PF04611	PF05185	PF05902	PF06816	PF07817	PF08243
PF00085	PF00467	PF00867	PF01239	PF01717	PF02199	PF02824	PF03416	PF03982	PF04614	PF05186	PF05903	PF06817	PF07818	PF08244
PF00087	PF00468	PF00868	PF01242	PF01722	PF02200	PF02825	PF03417	PF03983	PF04615	PF05187	PF05912	PF06818	PF07819	PF08245
PF00088	PF00472	PF00870	PF01243	PF01723	PF02201	PF02826	PF03422	PF03985	PF04617	PF05188	PF05916	PF06821	PF07822	PF08246
PF00090	PF00473	PF00875	PF01244	PF01725	PF02202	PF02827	PF03437	PF03986	PF04621	PF05189	PF05918	PF06824	PF07829	PF08248
PF00091	PF00474	PF00876	PF01245	PF01728	PF02204	PF02828	PF03439	PF03987	PF04622	PF05190	PF05922	PF06825	PF07830	PF08249
PF00093	PF00475	PF00879	PF01246	PF01729	PF02205	PF02833	PF03441	PF03990	PF04625	PF05191	PF05923	PF06827	PF07831	PF08250
PF00095	PF00476	PF00881	PF01247	PF01731	PF02206	PF02836	PF03443	PF03991	PF04626	PF05195	PF05924	PF06828	PF07834	PF08251
PF00096	PF00479	PF00884	PF01248	PF01733	PF02207	PF02840	PF03446	PF03992	PF04627	PF05196	PF05933	PF06831	PF07837	PF08252
PF00098	PF00480	PF00885	PF01249	PF01734	PF02208	PF02841	PF03447	PF03997	PF04628	PF05199	PF05937	PF06839	PF07842	PF08256
PF00107	PF00481	PF00886	PF01251	PF01740	PF02209	PF02843	PF03450	PF03998	PF04629	PF05202	PF05956	PF06858	PF07847	PF08257
PF00108	PF00484	PF00887	PF01253	PF01743	PF02212	PF02844	PF03451	PF03999	PF04636	PF05204	PF05958	PF06862	PF07850	PF08258
PF00112	PF00485	PF00889	PF01254	PF01746	PF02213	PF02845	PF03452	PF04001	PF04641	PF05205	PF05964	PF06870	PF07851	PF08259
PF00113	PF00487	PF00890	PF01255	PF01747	PF02218	PF02847	PF03453	PF04003	PF04643	PF05207	PF05965	PF06875	PF07856	PF08260
PF00116	PF00488	PF00891	PF01256	PF01748	PF02219	PF02852	PF03454	PF04004	PF04648	PF05208	PF05967	PF06881	PF07857	PF08261
PF00117	PF00489	PF00892	PF01257	PF01749	PF02221	PF02861	PF03455	PF04005	PF04652	PF05216	PF05971	PF06886	PF07859	PF08262
PF00121	PF00490	PF00895	PF01259	PF01751	PF02222	PF02862	PF03456	PF04006	PF04658	PF05222	PF05972	PF06888	PF07883	PF08265
PF00123	PF00491	PF00896	PF01261	PF01753	PF02223	PF02864	PF03460	PF04030	PF04664	PF05224	PF05978	PF06905	PF07884	PF08266
PF00129	PF00493	PF00899	PF01263	PF01754	PF02225	PF02865	PF03462	PF04031	PF04667	PF05225	PF05980	PF06907	PF07885	PF08267
PF00130	PF00494	PF00900	PF01264	PF01756	PF02229	PF02866	PF03463	PF04032	PF04668	PF05234	PF05995	PF06910	PF07894	PF08271
PF00131	PF00498	PF00903	PF01265	PF01757	PF02230	PF02867	PF03464	PF04034	PF04670	PF05236	PF05997	PF06916	PF07896	PF08279
PF00132	PF00499	PF00904	PF01267	PF01758	PF02233	PF02870	PF03465	PF04037	PF04675	PF05237	PF06001	PF06920	PF07904	PF08285
PF00136	PF00507	PF00909	PF01268	PF01759	PF02234	PF02872	PF03467	PF04042	PF04676	PF05238	PF06008	PF06936	PF07912	PF08286
PF00137	PF00510	PF00916	PF01269	PF01761	PF02237	PF02874	PF03477	PF04044	PF04677	PF05240	PF06009	PF06937	PF07915	PF08287
PF00141	PF00512	PF00917	PF01271	PF01762	PF02238	PF02875	PF03481	PF04045	PF04679	PF05241	PF06010	PF06941	PF07923	PF08288
PF00144	PF00521	PF00919	PF01273	PF01765	PF02244	PF02877	PF03483	PF04046	PF04680	PF05242	PF06011	PF06951	PF07926	PF08292

PF00145	PF00525	PF00920	PF01274	PF01770	PF02245	PF02879	PF03484	PF04047	PF04683	PF05250	PF06012	PF06954	PF07928	PF08293
PF00146	PF00531	PF00923	PF01275	PF01773	PF02251	PF02880	PF03485	PF04048	PF04691	PF05251	PF06017	PF06957	PF07933	PF08295
PF00152	PF00538	PF00925	PF01279	PF01775	PF02252	PF02881	PF03487	PF04049	PF04692	PF05253	PF06020	PF06959	PF07934	PF08296
PF00153	PF00542	PF00926	PF01280	PF01776	PF02257	PF02882	PF03489	PF04050	PF04695	PF05254	PF06025	PF06963	PF07936	PF08297
PF00157	PF00544	PF00928	PF01281	PF01777	PF02259	PF02883	PF03491	PF04051	PF04696	PF05255	PF06026	PF06964	PF07938	PF08302
PF00159	PF00545	PF00929	PF01282	PF01778	PF02260	PF02885	PF03493	PF04053	PF04697	PF05270	PF06031	PF06968	PF07941	PF08303
PF00160	PF00547	PF00930	PF01283	PF01779	PF02262	PF02887	PF03494	PF04054	PF04698	PF05276	PF06046	PF06970	PF07942	PF08304
PF00162	PF00549	PF00931	PF01284	PF01780	PF02265	PF02888	PF03501	PF04055	PF04699	PF05277	PF06049	PF06978	PF07945	PF08311
PF00163	PF00551	PF00932	PF01285	PF01781	PF02267	PF02889	PF03508	PF04056	PF04704	PF05279	PF06052	PF06984	PF07946	PF08312
PF00166	PF00552	PF00933	PF01286	PF01783	PF02268	PF02891	PF03509	PF04057	PF04706	PF05281	PF06058	PF06985	PF07954	PF08313
PF00171	PF00554	PF00935	PF01287	PF01784	PF02269	PF02892	PF03511	PF04059	PF04707	PF05282	PF06060	PF06990	PF07955	PF08315
PF00172	PF00560	PF00940	PF01288	PF01786	PF02270	PF02894	PF03516	PF04061	PF04709	PF05283	PF06068	PF06991	PF07956	PF08316
PF00173	PF00562	PF00941	PF01290	PF01791	PF02271	PF02897	PF03517	PF04062	PF04710	PF05285	PF06071	PF06994	PF07957	PF08317
PF00174	PF00564	PF00953	PF01291	PF01793	PF02272	PF02898	PF03520	PF04063	PF04711	PF05287	PF06079	PF07001	PF07958	PF08318
PF00177	PF00566	PF00956	PF01293	PF01794	PF02274	PF02902	PF03521	PF04064	PF04714	PF05291	PF06083	PF07002	PF07959	PF08319
PF00178	PF00568	PF00958	PF01300	PF01798	PF02275	PF02906	PF03522	PF04065	PF04715	PF05292	PF06087	PF07010	PF07960	PF08320
PF00179	PF00569	PF00961	PF01302	PF01799	PF02284	PF02910	PF03523	PF04068	PF04716	PF05296	PF06093	PF07019	PF07961	PF08321
PF00180	PF00570	PF00962	PF01303	PF01803	PF02285	PF02911	PF03525	PF04072	PF04718	PF05298	PF06105	PF07034	PF07962	PF08322
PF00181	PF00572	PF00969	PF01305	PF01805	PF02290	PF02913	PF03528	PF04078	PF04719	PF05300	PF06119	PF07035	PF07964	PF08325
PF00184	PF00573	PF00975	PF01315	PF01808	PF02291	PF02919	PF03529	PF04080	PF04722	PF05303	PF06121	PF07054	PF07965	PF08326
PF00185	PF00574	PF00977	PF01323	PF01812	PF02292	PF02920	PF03530	PF04081	PF04724	PF05316	PF06127	PF07061	PF07966	PF08327
PF00186	PF00575	PF00984	PF01327	PF01813	PF02296	PF02921	PF03531	PF04083	PF04727	PF05324	PF06140	PF07062	PF07967	PF08332
PF00187	PF00576	PF00988	PF01328	PF01814	PF02297	PF02922	PF03533	PF04084	PF04729	PF05326	PF06148	PF07064	PF07970	PF08333
PF00188	PF00578	PF00993	PF01329	PF01817	PF02301	PF02928	PF03535	PF04086	PF04731	PF05327	PF06151	PF07074	PF07973	PF08336
PF00189	PF00579	PF00994	PF01331	PF01822	PF02310	PF02929	PF03540	PF04088	PF04732	PF05328	PF06172	PF07081	PF07974	PF08337
PF00191	PF00580	PF00996	PF01335	PF01823	PF02312	PF02933	PF03542	PF04089	PF04733	PF05337	PF06179	PF07084	PF07975	PF08344
PF00193	PF00581	PF00999	PF01336	PF01827	PF02318	PF02934	PF03547	PF04090	PF04736	PF05345	PF06189	PF07093	PF07976	PF08351
PF00198	PF00582	PF01000	PF01341	PF01828	PF02319	PF02935	PF03556	PF04091	PF04739	PF05346	PF06201	PF07106	PF07978	PF08354
PF00199	PF00584	PF01007	PF01342	PF01833	PF02320	PF02937	PF03561	PF04095	PF04749	PF05347	PF06202	PF07111	PF07985	PF08355
PF00202	PF00585	PF01008	PF01347	PF01835	PF02330	PF02939	PF03567	PF04096	PF04750	PF05348	PF06209	PF07140	PF07986	PF08356
PF00203	PF00587	PF01011	PF01348	PF01839	PF02337	PF02940	PF03568	PF04097	PF04752	PF05349	PF06211	PF07142	PF07988	PF08357
PF00204	PF00588	PF01012	PF01351	PF01840	PF02338	PF02944	PF03571	PF04098	PF04756	PF05350	PF06212	PF07145	PF07989	PF08365

PF00206	PF00590	PF01014	PF01353	PF01841	PF02344	PF02946	PF03572	PF04099	PF04757	PF05351	PF06214	PF07147	PF07991	PF08366
PF00207	PF00591	PF01015	PF01354	PF01842	PF02345	PF02947	PF03575	PF04100	PF04758	PF05353	PF06218	PF07149	PF07993	PF08367
PF00208	PF00594	PF01016	PF01363	PF01843	PF02347	PF02948	PF03577	PF04101	PF04760	PF05355	PF06220	PF07156	PF07994	PF08368
PF00209	PF00596	PF01017	PF01365	PF01844	PF02348	PF02949	PF03583	PF04103	PF04762	PF05362	PF06221	PF07159	PF08005	PF08373
PF00210	PF00605	PF01018	PF01368	PF01846	PF02350	PF02961	PF03587	PF04104	PF04768	PF05365	PF06229	PF07163	PF08015	PF08374
PF00211	PF00609	PF01020	PF01369	PF01847	PF02351	PF02965	PF03595	PF04106	PF04774	PF05366	PF06237	PF07165	PF08016	PF08375
PF00213	PF00610	PF01021	PF01372	PF01852	PF02354	PF02966	PF03600	PF04109	PF04775	PF05374	PF06248	PF07169	PF08022	PF08377
PF00214	PF00612	PF01023	PF01379	PF01853	PF02358	PF02969	PF03604	PF04110	PF04777	PF05378	PF06268	PF07175	PF08023	PF08383
PF00215	PF00613	PF01025	PF01380	PF01857	PF02359	PF02970	PF03615	PF04111	PF04790	PF05383	PF06292	PF07177	PF08024	PF08384
PF00218	PF00614	PF01026	PF01382	PF01858	PF02364	PF02971	PF03619	PF04112	PF04791	PF05386	PF06294	PF07200	PF08025	PF08385
PF00219	PF00616	PF01027	PF01384	PF01866	PF02366	PF02973	PF03623	PF04113	PF04795	PF05387	PF06297	PF07202	PF08026	PF08389
PF00220	PF00617	PF01028	PF01387	PF01868	PF02372	PF02978	PF03630	PF04114	PF04799	PF05388	PF06309	PF07225	PF08029	PF08390
PF00221	PF00618	PF01033	PF01388	PF01869	PF02374	PF02982	PF03632	PF04115	PF04800	PF05390	PF06311	PF07228	PF08030	PF08393
PF00224	PF00621	PF01035	PF01390	PF01872	PF02376	PF02985	PF03633	PF04116	PF04801	PF05391	PF06312	PF07247	PF08032	PF08397
PF00225	PF00623	PF01036	PF01392	PF01873	PF02377	PF02987	PF03635	PF04117	PF04802	PF05392	PF06313	PF07249	PF08033	PF08399
PF00230	PF00624	PF01039	PF01393	PF01875	PF02383	PF02990	PF03636	PF04118	PF04803	PF05395	PF06314	PF07253	PF08034	PF08403
PF00231	PF00627	PF01040	PF01394	PF01876	PF02386	PF02991	PF03637	PF04119	PF04810	PF05397	PF06320	PF07260	PF08035	PF08407
PF00232	PF00629	PF01042	PF01395	PF01883	PF02390	PF02994	PF03639	PF04120	PF04811	PF05399	PF06324	PF07263	PF08036	PF08412
PF00235	PF00632	PF01048	PF01396	PF01885	PF02391	PF02996	PF03641	PF04121	PF04812	PF05404	PF06328	PF07264	PF08037	PF08416
PF00236	PF00635	PF01055	PF01397	PF01894	PF02403	PF03002	PF03643	PF04124	PF04813	PF05405	PF06330	PF07276	PF08038	PF08418
PF00237	PF00636	PF01058	PF01398	PF01896	PF02404	PF03006	PF03645	PF04127	PF04814	PF05406	PF06331	PF07281	PF08039	PF08420
PF00238	PF00638	PF01059	PF01401	PF01900	PF02410	PF03016	PF03647	PF04128	PF04815	PF05418	PF06333	PF07286	PF08040	PF08424
PF00241	PF00639	PF01060	PF01403	PF01907	PF02412	PF03020	PF03648	PF04129	PF04818	PF05427	PF06337	PF07292	PF08045	PF08426
PF00243	PF00642	PF01061	PF01404	PF01910	PF02422	PF03024	PF03651	PF04130	PF04819	PF05428	PF06342	PF07294	PF08058	PF08429
PF00244	PF00643	PF01063	PF01410	PF01912	PF02423	PF03027	PF03657	PF04132	PF04821	PF05433	PF06345	PF07303	PF08059	PF08430
PF00245	PF00644	PF01064	PF01411	PF01915	PF02425	PF03029	PF03659	PF04133	PF04822	PF05434	PF06350	PF07324	PF08060	PF08432
PF00246	PF00645	PF01066	PF01412	PF01916	PF02434	PF03031	PF03660	PF04135	PF04824	PF05436	PF06357	PF07327	PF08061	PF08434
PF00248	PF00646	PF01067	PF01413	PF01918	PF02436	PF03032	PF03661	PF04136	PF04825	PF05438	PF06365	PF07334	PF08062	PF08438
PF00250	PF00647	PF01068	PF01414	PF01922	PF02437	PF03033	PF03663	PF04137	PF04831	PF05439	PF06367	PF07347	PF08063	PF08441
PF00251	PF00649	PF01070	PF01415	PF01923	PF02441	PF03034	PF03664	PF04139	PF04836	PF05450	PF06369	PF07353	PF08064	PF08442
PF00252	PF00653	PF01071	PF01416	PF01926	PF02450	PF03036	PF03665	PF04140	PF04840	PF05453	PF06372	PF07354	PF08065	PF08448
PF00253	PF00654	PF01073	PF01417	PF01928	PF02453	PF03039	PF03666	PF04144	PF04841	PF05454	PF06373	PF07359	PF08066	PF08449

PF00255	PF00656	PF01074	PF01419	PF01929	PF02460	PF03045	PF03669	PF04145	PF04845	PF05456	PF06374	PF07365	PF08067	PF08451
PF00256	PF00657	PF01077	PF01421	PF01933	PF02466	PF03051	PF03670	PF04146	PF04847	PF05458	PF06375	PF07366	PF08068	PF08454
PF00258	PF00658	PF01079	PF01422	PF01937	PF02469	PF03055	PF03671	PF04147	PF04855	PF05460	PF06384	PF07367	PF08069	PF08466
PF00260	PF00659	PF01080	PF01424	PF01938	PF02485	PF03056	PF03674	PF04148	PF04856	PF05461	PF06387	PF07393	PF08070	PF08473
PF00262	PF00662	PF01082	PF01425	PF01946	PF02487	PF03061	PF03676	PF04152	PF04858	PF05463	PF06391	PF07400	PF08071	PF08474
PF00265	PF00666	PF01083	PF01426	PF01958	PF02492	PF03062	PF03690	PF04153	PF04868	PF05466	PF06393	PF07406	PF08072	PF08477
PF00268	PF00670	PF01084	PF01428	PF01960	PF02493	PF03068	PF03694	PF04157	PF04869	PF05469	PF06394	PF07412	PF08073	PF08487
PF00272	PF00673	PF01085	PF01432	PF01963	PF02494	PF03070	PF03700	PF04158	PF04871	PF05470	PF06395	PF07421	PF08074	PF08490
PF00273	PF00676	PF01086	PF01434	PF01965	PF02513	PF03073	PF03712	PF04163	PF04874	PF05476	PF06396	PF07425	PF08075	PF08491
PF00274	PF00677	PF01087	PF01435	PF01966	PF02515	PF03074	PF03715	PF04176	PF04882	PF05477	PF06398	PF07440	PF08079	PF08492
PF00275	PF00679	PF01088	PF01436	PF01967	PF02516	PF03081	PF03718	PF04177	PF04889	PF05478	PF06399	PF07442	PF08080	PF08493
PF00278	PF00683	PF01090	PF01437	PF01968	PF02517	PF03082	PF03719	PF04178	PF04892	PF05482	PF06400	PF07443	PF08081	PF08498
PF00280	PF00684	PF01091	PF01439	PF01974	PF02520	PF03088	PF03720	PF04179	PF04893	PF05483	PF06401	PF07448	PF08082	PF08499
PF00281	PF00685	PF01093	PF01448	PF01975	PF02525	PF03089	PF03721	PF04180	PF04898	PF05485	PF06403	PF07452	PF08083	PF08501
PF00285	PF00686	PF01095	PF01450	PF01981	PF02535	PF03091	PF03722	PF04181	PF04900	PF05486	PF06417	PF07458	PF08084	PF08502
PF00287	PF00687	PF01096	PF01451	PF01984	PF02536	PF03095	PF03723	PF04182	PF04901	PF05492	PF06418	PF07464	PF08086	PF08504
PF00288	PF00689	PF01101	PF01459	PF01988	PF02538	PF03096	PF03725	PF04185	PF04902	PF05493	PF06419	PF07469	PF08088	PF08506
PF00290	PF00690	PF01103	PF01462	PF01991	PF02541	PF03097	PF03726	PF04188	PF04904	PF05495	PF06420	PF07473	PF08089	PF08508
PF00291	PF00692	PF01105	PF01463	PF01992	PF02544	PF03099	PF03727	PF04189	PF04905	PF05499	PF06421	PF07474	PF08091	PF08509
PF00292	PF00694	PF01106	PF01464	PF01997	PF02545	PF03102	PF03730	PF04191	PF04908	PF05502	PF06423	PF07477	PF08092	PF08510
PF00293	PF00696	PF01109	PF01465	PF02005	PF02548	PF03104	PF03731	PF04192	PF04909	PF05507	PF06424	PF07479	PF08093	PF08513
PF00297	PF00699	PF01110	PF01466	PF02008	PF02550	PF03105	PF03733	PF04193	PF04910	PF05508	PF06425	PF07488	PF08094	PF08514
PF00298	PF00706	PF01111	PF01467	PF02012	PF02551	PF03109	PF03735	PF04194	PF04911	PF05510	PF06427	PF07491	PF08095	PF08515
PF00300	PF00707	PF01112	PF01469	PF02013	PF02567	PF03124	PF03736	PF04201	PF04912	PF05517	PF06428	PF07492	PF08096	PF08516
PF00303	PF00708	PF01114	PF01470	PF02014	PF02569	PF03126	PF03737	PF04209	PF04921	PF05522	PF06432	PF07493	PF08097	PF08517
PF00304	PF00709	PF01115	PF01472	PF02015	PF02578	PF03127	PF03746	PF04212	PF04923	PF05529	PF06441	PF07496	PF08098	PF08518
PF00305	PF00710	PF01116	PF01474	PF02017	PF02580	PF03128	PF03747	PF04218	PF04925	PF05546	PF06446	PF07500	PF08099	PF08519
PF00306	PF00711	PF01118	PF01477	PF02019	PF02581	PF03129	PF03758	PF04230	PF04926	PF05571	PF06448	PF07502	PF08101	PF08520
PF00307	PF00713	PF01119	PF01479	PF02020	PF02582	PF03130	PF03762	PF04253	PF04930	PF05572	PF06451	PF07519	PF08102	PF08523
PF00313	PF00714	PF01120	PF01480	PF02022	PF02585	PF03131	PF03765	PF04258	PF04934	PF05577	PF06456	PF07521	PF08103	PF08526
PF00317	PF00715	PF01121	PF01482	PF02025	PF02586	PF03133	PF03768	PF04272	PF04935	PF05586	PF06457	PF07522	PF08104	PF08527
PF00318	PF00717	PF01122	PF01483	PF02026	PF02594	PF03134	PF03769	PF04275	PF04938	PF05587	PF06459	PF07524	PF08105	PF08528

PF00319	PF00719	PF01124	PF01484	PF02035	PF02597	PF03145	PF03770	PF04280	PF04939	PF05600	PF06461	PF07525	PF08106	PF08534
PF00320	PF00723	PF01125	PF01485	PF02036	PF02598	PF03146	PF03778	PF04281	PF04950	PF05604	PF06462	PF07527	PF08108	PF08537
PF00323	PF00724	PF01126	PF01487	PF02037	PF02602	PF03147	PF03781	PF04299	PF04952	PF05609	PF06463	PF07528	PF08110	PF08540
PF00324	PF00727	PF01127	PF01491	PF02039	PF02607	PF03148	PF03782	PF04300	PF04959	PF05622	PF06464	PF07529	PF08111	PF08542
PF00327	PF00728	PF01129	PF01493	PF02044	PF02617	PF03149	PF03792	PF04318	PF04960	PF05637	PF06465	PF07530	PF08114	PF08544
PF00328	PF00731	PF01130	PF01496	PF02045	PF02626	PF03151	PF03798	PF04324	PF04961	PF05640	PF06466	PF07531	PF08115	PF08546
PF00329	PF00733	PF01131	PF01498	PF02046	PF02627	PF03152	PF03800	PF04325	PF04968	PF05641	PF06467	PF07533	PF08116	
PF00330	PF00734	PF01133	PF01502	PF02058	PF02628	PF03153	PF03801	PF04366	PF04969	PF05645	PF06468	PF07535	PF08117	
PF00331	PF00735	PF01134	PF01503	PF02059	PF02629	PF03155	PF03803	PF04370	PF04970	PF05648	PF06469	PF07539	PF08119	
PF00333	PF00736	PF01135	PF01504	PF02064	PF02630	PF03159	PF03807	PF04376	PF04979	PF05649	PF06470	PF07540	PF08120	
PF00334	PF00742	PF01139	PF01505	PF02066	PF02636	PF03160	PF03810	PF04377	PF04981	PF05653	PF06472	PF07541	PF08121	
PF00335	PF00743	PF01140	PF01507	PF02067	PF02637	PF03161	PF03813	PF04379	PF04983	PF05669	PF06479	PF07542	PF08122	
PF00338	PF00748	PF01142	PF01509	PF02071	PF02656	PF03164	PF03815	PF04382	PF04987	PF05670	PF06480	PF07543	PF08123	
PF00342	PF00749	PF01144	PF01512	PF02076	PF02666	PF03165	PF03820	PF04387	PF04988	PF05671	PF06482	PF07545	PF08125	
PF00343	PF00751	PF01146	PF01513	PF02077	PF02671	PF03166	PF03821	PF04388	PF04990	PF05676	PF06487	PF07546	PF08127	
PF00344	PF00752	PF01148	PF01521	PF02078	PF02676	PF03167	PF03823	PF04406	PF04992	PF05679	PF06495	PF07557	PF08131	
PF00346	PF00755	PF01150	PF01522	PF02079	PF02678	PF03169	PF03824	PF04408	PF05000	PF05680	PF06512	PF07558	PF08138	
PF00347	PF00756	PF01151	PF01529	PF02083	PF02682	PF03171	PF03826	PF04410	PF05001	PF05682	PF06519	PF07562	PF08140	
PF00348	PF00758	PF01153	PF01530	PF02084	PF02688	PF03172	PF03827	PF04418	PF05002	PF05693	PF06524	PF07565	PF08142	
PF00350	PF00762	PF01154	PF01531	PF02087	PF02696	PF03177	PF03828	PF04419	PF05004	PF05700	PF06534	PF07569	PF08143	
PF00351	PF00763	PF01156	PF01532	PF02088	PF02705	PF03178	PF03832	PF04420	PF05005	PF05706	PF06535	PF07571	PF08144	
PF00352	PF00764	PF01157	PF01534	PF02089	PF02709	PF03179	PF03834	PF04421	PF05007	PF05712	PF06540	PF07572	PF08145	
PF00354	PF00766	PF01158	PF01536	PF02093	PF02714	PF03184	PF03835	PF04423	PF05010	PF05715	PF06544	PF07573	PF08146	
PF00355	PF00769	PF01159	PF01541	PF02096	PF02724	PF03185	PF03836	PF04424	PF05014	PF05719	PF06546	PF07575	PF08147	
PF00357	PF00773	PF01160	PF01544	PF02098	PF02728	PF03188	PF03839	PF04425	PF05019	PF05721	PF06553	PF07576	PF08148	
PF00362	PF00774	PF01161	PF01545	PF02100	PF02729	PF03190	PF03847	PF04426	PF05020	PF05722	PF06554	PF07593	PF08149	
PF00364	PF00777	PF01162	PF01551	PF02101	PF02731	PF03194	PF03849	PF04427	PF05021	PF05724	PF06566	PF07647	PF08150	
PF00365	PF00778	PF01163	PF01554	PF02102	PF02732	PF03198	PF03850	PF04428	PF05022	PF05726	PF06567	PF07649	PF08151	
PF00366	PF00779	PF01166	PF01557	PF02104	PF02733	PF03199	PF03851	PF04433	PF05023	PF05729	PF06573	PF07650	PF08152	
PF00368	PF00780	PF01167	PF01562	PF02106	PF02734	PF03200	PF03853	PF04434	PF05024	PF05730	PF06581	PF07651	PF08153	
PF00372	PF00784	PF01168	PF01564	PF02109	PF02735	PF03208	PF03855	PF04437	PF05026	PF05731	PF06583	PF07657	PF08154	
PF00375	PF00786	PF01169	PF01565	PF02110	PF02737	PF03215	PF03856	PF04438	PF05028	PF05735	PF06584	PF07662	PF08155	

PF00377	PF00787	PF01170	PF01569	PF02112	PF02738	PF03224	PF03858	PF04440	PF05030	PF05741	PF06588	PF07670	PF08156	
PF00378	PF00790	PF01171	PF01575	PF02114	PF02740	PF03226	PF03867	PF04442	PF05032	PF05742	PF06602	PF07677	PF08157	
PF00382	PF00791	PF01172	PF01581	PF02115	PF02743	PF03227	PF03868	PF04446	PF05033	PF05743	PF06608	PF07678	PF08158	
PF00383	PF00792	PF01174	PF01583	PF02116	PF02744	PF03232	PF03870	PF04468	PF05038	PF05746	PF06614	PF07683	PF08159	
PF00384	PF00793	PF01175	PF01586	PF02121	PF02747	PF03234	PF03871	PF04478	PF05039	PF05753	PF06621	PF07684	PF08160	
PF00386	PF00795	PF01176	PF01588	PF02126	PF02749	PF03239	PF03874	PF04479	PF05040	PF05756	PF06623	PF07687	PF08161	
PF00388	PF00797	PF01177	PF01591	PF02127	PF02750	PF03247	PF03875	PF04487	PF05041	PF05760	PF06625	PF07691	PF08162	
PF00390	PF00800	PF01179	PF01592	PF02130	PF02751	PF03248	PF03876	PF04488	PF05046	PF05761	PF06628	PF07699	PF08163	
PF00393	PF00801	PF01180	PF01595	PF02133	PF02755	PF03250	PF03878	PF04493	PF05047	PF05764	PF06631	PF07701	PF08164	
PF00394	PF00804	PF01183	PF01596	PF02134	PF02757	PF03253	PF03879	PF04494	PF05051	PF05769	PF06632	PF07703	PF08165	
PF00396	PF00806	PF01184	PF01597	PF02135	PF02758	PF03256	PF03893	PF04495	PF05053	PF05777	PF06637	PF07706	PF08168	
PF00397	PF00808	PF01185	PF01598	PF02137	PF02759	PF03259	PF03894	PF04499	PF05057	PF05778	PF06638	PF07710	PF08169	