UNIVERSITÉ DE GENÈVE

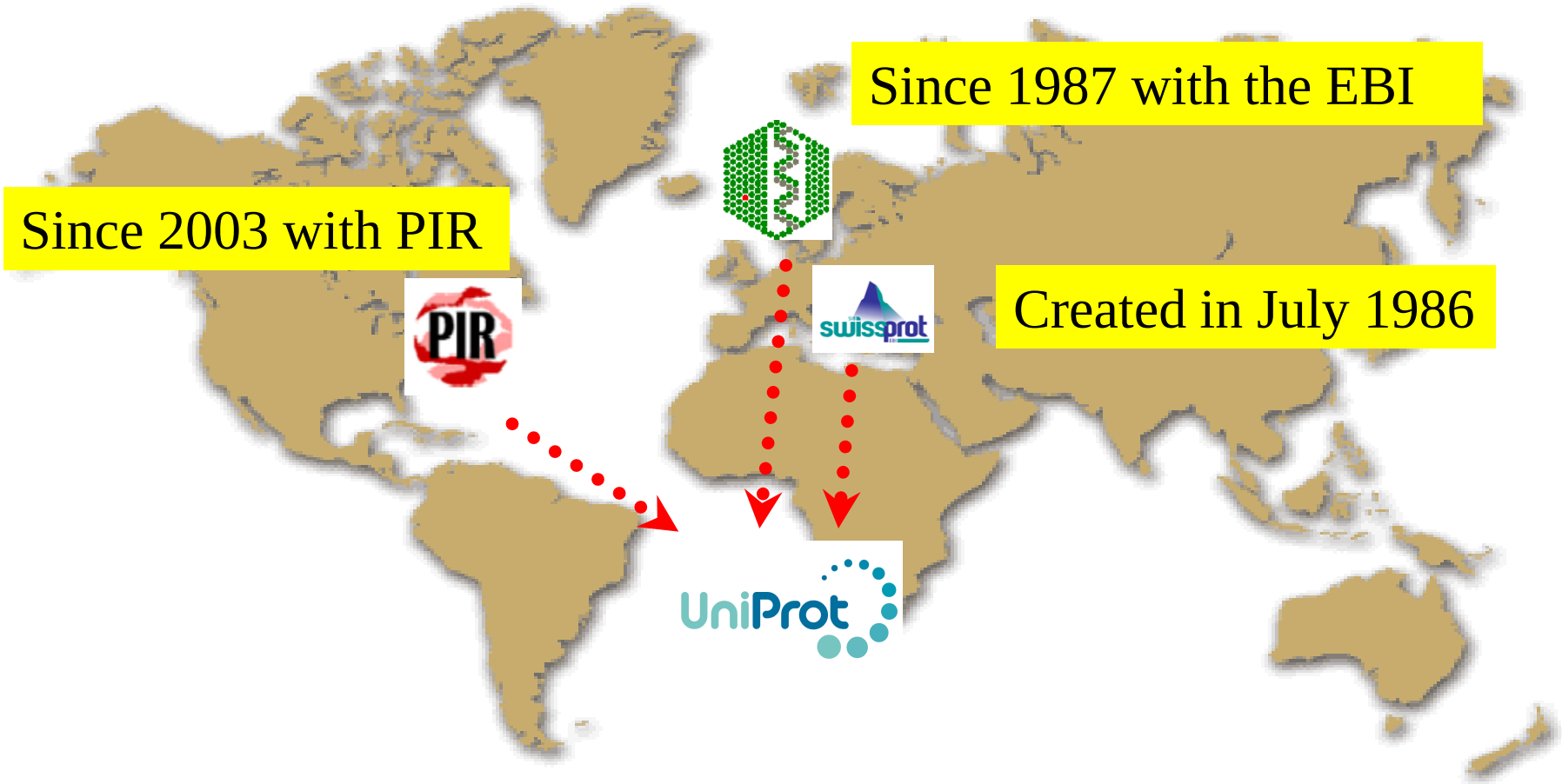
SIB Swiss Institute of Bioinformatics

neXtProt: a new human-centric protein knowledge resource

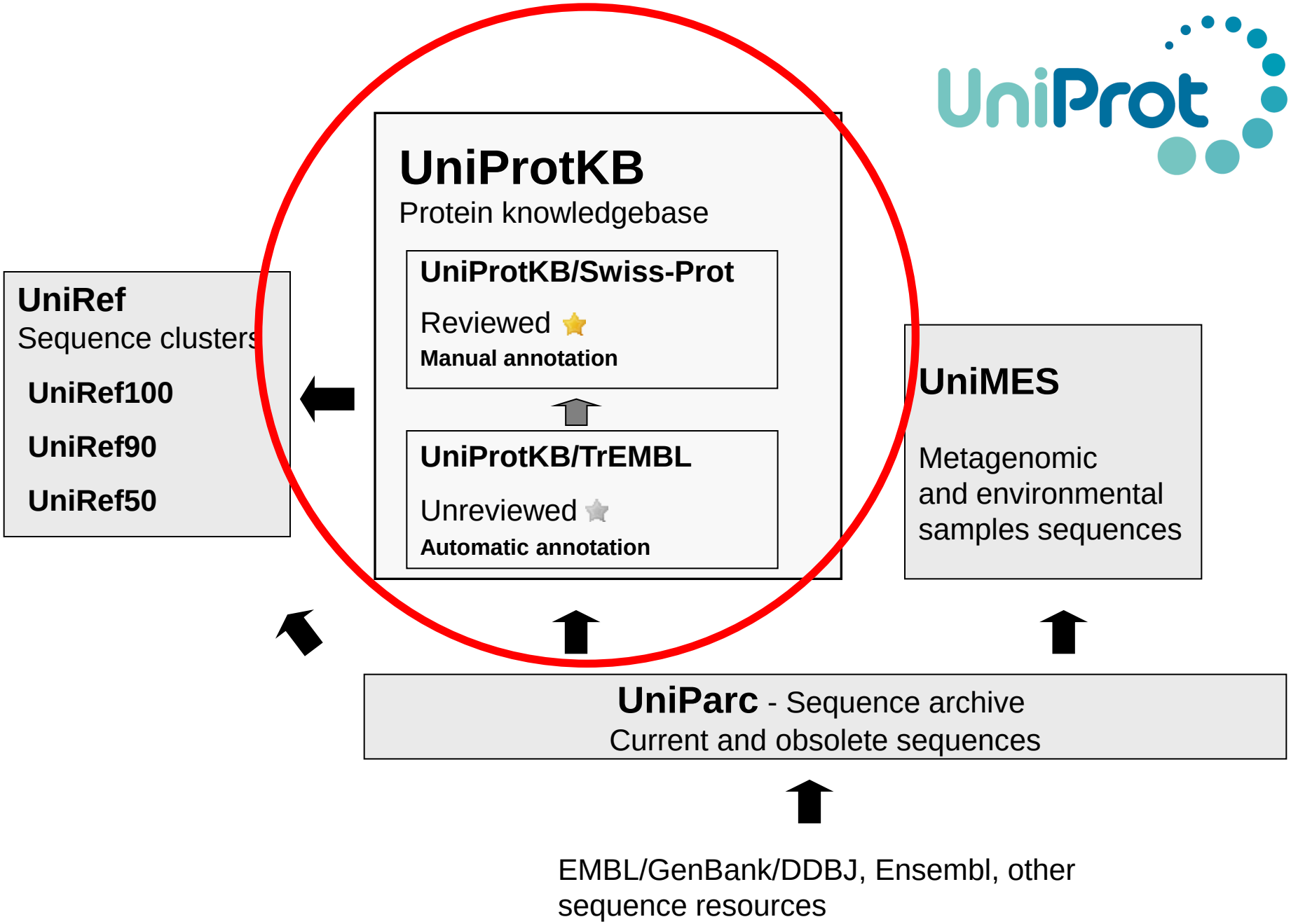
Amos Bairoch

July 14, 2010

From Swiss-Prot to UniProt



EBI, PIR and SIB together form the UniProt consortium


UniProt
UniProtKB
Protein knowledgebase
UniProtKB/Swiss-Prot
Reviewed
Manual annotation
UniProtKB/TrEMBL
Unreviewed
Automatic annotation
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniMES
Metagenomic
and environmental
samples sequences
UniParc - Sequence archive
Current and obsolete sequences
EMBL/GenBank/DDBJ, Ensembl, other
sequence resources

4th SIENA MEETING

FROM GENOME TO PROTEOME:
KNOWLEDGE ACQUISITION AND REPRESENTATION

Sept. 4-7, 2000, Siena, Italy

Almost 10 years ago, at the 4th Siena meeting, we proposed to annotate in Swiss-Prot all the human proteins

178 Review TRENDS in Biotechnology Vol.19 No.5 May 2001

The human proteomics initiative (HPI)

8TH SIENA MEETING

FROM GENOME TO PROTEOME:

INTEGRATION AND PROTEOME COMPLETION

Siena, Italy, August 31st- September 4th, 2008

Auditorium Giurisprudenza e Scienze Politiche

ProteoMonitor

THE GLOBAL NEWSWEEKLY OF PROTEOMICS TECHNOLOGY

UniProt Releases 'Complete' Set of 20K Human Proteins at Siena Meeting

[September 4, 2008]

A ‘complete’ set of annotated human proteins

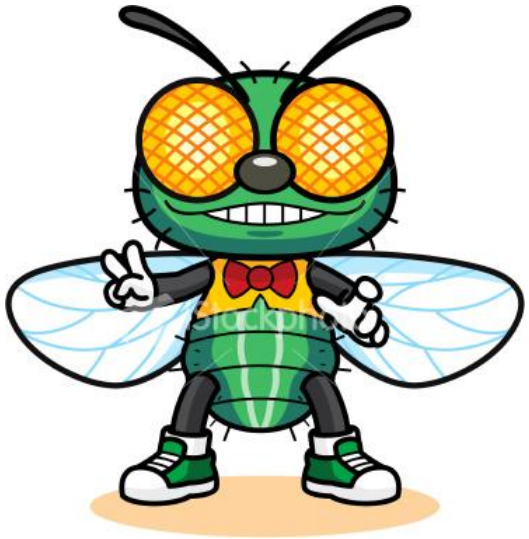
- In September 2008, we had annotated **20’330** human protein entries in UniProtKB/Swiss-Prot;
- They originate from about **20’400** protein-coding genes;
- Why ‘about’?
  - There are sets of genes that encode for identical proteins (example: 14 genes code for histone H4);
  - There are genes that codes for two or more proteins that have nothing in common in term of their sequence (bicistronic or alternative splicing);
  - There are some other weird cases!
- The precise definition of what is a gene is dependent on who is using/making that definition.

What do we mean by complete?

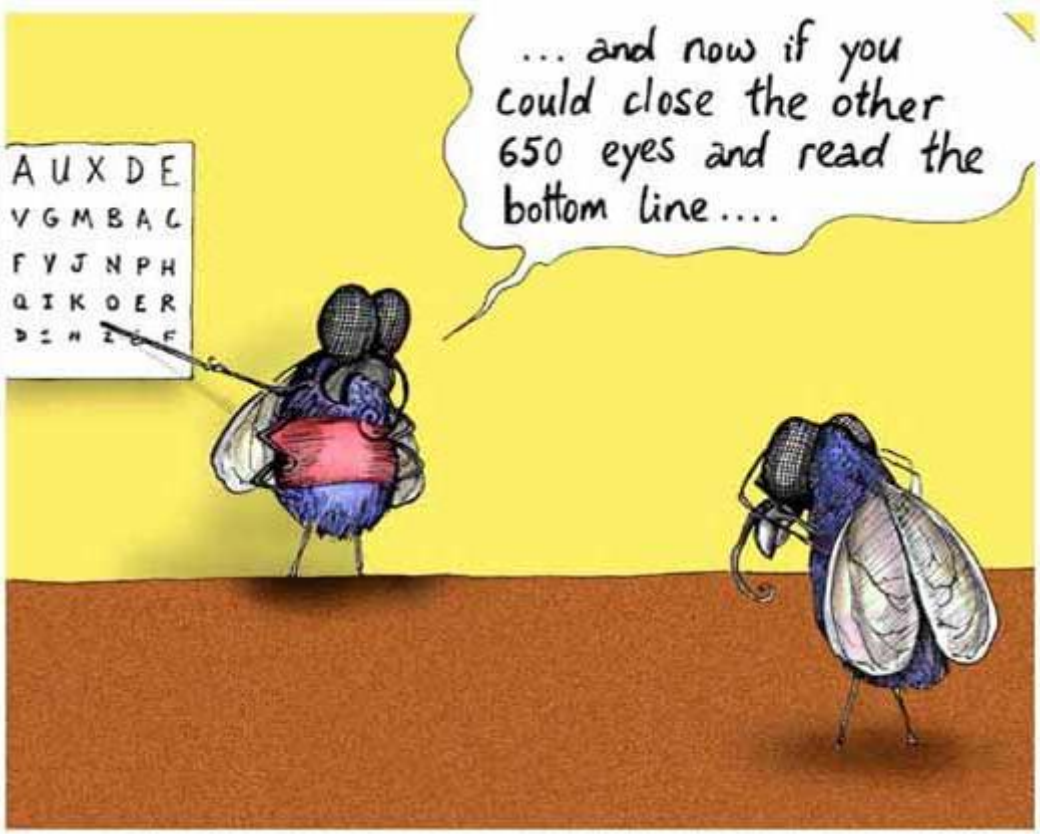
- We annotated all the protein-coding genes with an HGNC gene symbol;
- + All the predicted Ensembl genes validated by a variety of studies (including that of Michele Clamp and colleagues [PNAS 104:19428-33(2007)])
- + All those in the CCDS list;
- + All those referenced in OMIM;
- + All the ‘valid’ proteins from a series of full length cDNA projects (CGI, SPDI, Kasuza, etc.);
- + Anything else that seemed real and that annotators encountered while reading papers.

Since...

- Since the beginning of 2009, we have added 85 «new» sequences, but we have «deleted» 108 proteins;
- Our gut feeling is that we are slowly but inexorably creeping toward slightly under 20’000 human-protein coding genes.



So do not feel bad if Drosophila has only slightly less genes that we have



And if you think about it:

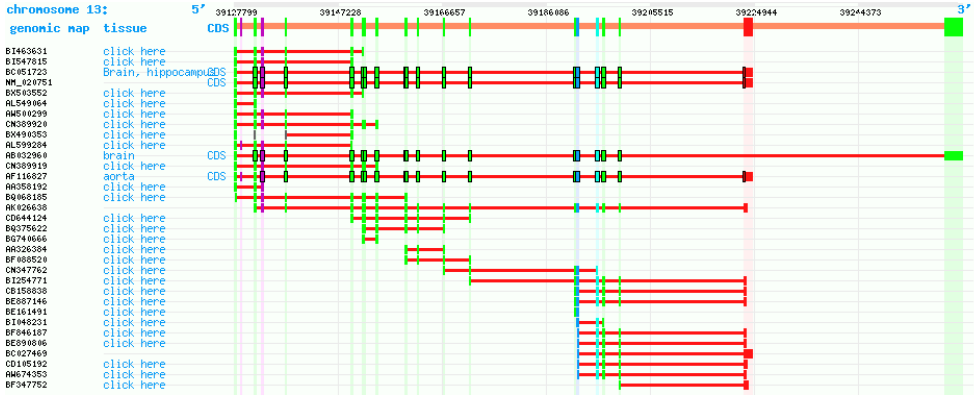
we can't fly,

we can't walk on the ceiling, and

we only have two eyes

Alternative isoforms

- Produced by alternative splicing, promoter usage or initiation;
- Currently we have **14’500** additional isoforms in about **7’500** entries;
- This means that **36%** of the protein-coding genes are already annotated to code for at least 2 different protein sequences;
- We estimate (based on an in-depth analysis of genes encoded on chromosome 13) that this number will rise above **60**% and the average number of isoforms to 3;
- This mean that we can already estimate that there is probably about **50’000** different human proteins that are produced by as many (or even more) transcripts.



This is going to be a long term problem

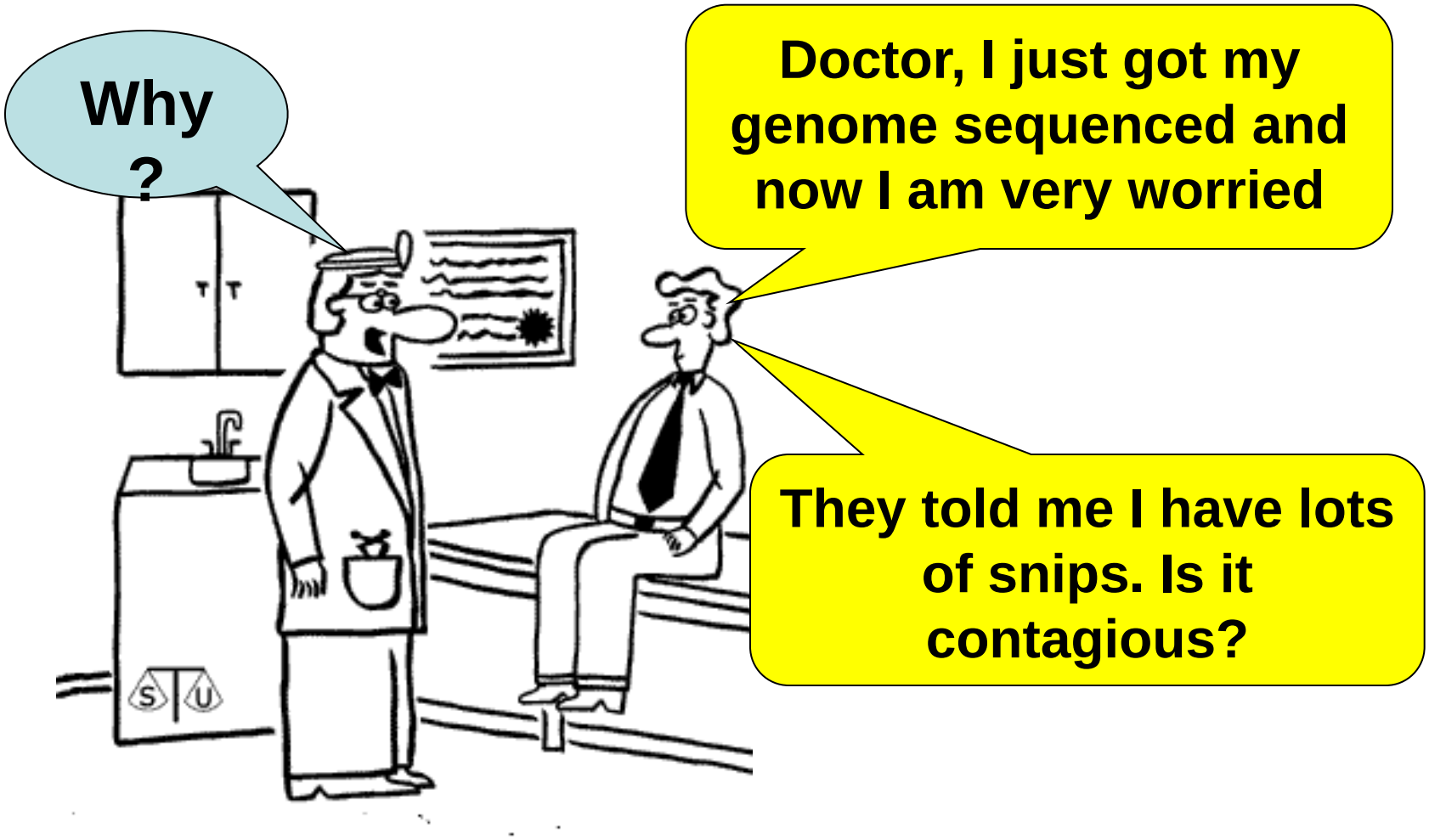
- Many isoforms are probably either not expressed or in tiny amount or in only restricted cell types;
- In term of proteomics:
  - It is not going to be easy to find prototypic signatures for all isoforms;
  - We probably need to specifically target a part of the identification effort toward the goal of establishing a complete catalog of the various expressed isoforms;
- And in term of annotation, we need to speed up the effort.

Sequence variants

- We have information concerning about **62’000** SAP (single amino-acid polymorphisms);
- **23’000** are linked to diseases. This information is mined from the literature and from disease-specific databases;
- This means that, excluding disease variants, there is already an average of 2 SAPs per protein;
- The ‘non-disease’ variants are obtained from a variety of sources (HAPMAP, NIEHS-SNPs, etc);
- They will increasingly come from whole human genome sequencing efforts (1’000 genomes, etc).

Caveat about variants

- The «canonical» human-genome derived sequence is an artefact;
- Some reported variants represent in fact the «majority» sequence;
- But we need to take into account that there is no such thing as an «average» human proteome!



Getting information on sequence variation is one thing, making sense of it, is something else

Post-translational modifications

- We have about **75’000** annotated PTMs;
- Only half of them have been experimentally obtained;
- The rest are predicted or inferred from experiments done in other species;
- We are just looking at the tip of the iceberg. But proteomics studies are starting to address this issue seriously;
- If we make a very modest estimate of 5 different PTMs per protein and that they may be independently regulated, you already get a **100x** increase in the number of protein species in our body (to a total of **5 million**).

The PTM world is still largely uncharted

(3R)-3-hydroxyasparagine, (3R)-3-hydroxyaspartate, (3S)-3-hydroxyasparagine, 1'-histidyl-3'-tyrosine, 1-thioglycine, 2',4',5'-topaquinone, 2,3-didehydroalanine, 3'-(S-cysteinyl)-tyrosine, 3-hydroxyproline, 3-oxoalanine, 4-amino-3-isothiazolidinone serine, 4-carboxyglutamate, 4-hydroxyproline, 5-glutamyl, 5-glutamyl glycerylphosphorylethanolamine, 5-hydroxylysine, 5-imidazolinone, ADP-ribosylasparagine, ADP-ribosylcysteine, ADP-ribosylserine, Allysine, Arginine amide, Asparagine amide, Aspartate 1-(chondroitin 4-sulfate)-ester, Asymmetric dimethylarginine, Beta-decarboxylated aspartate, Cholesterol glycine ester, Citrulline, Cysteine methyl ester, Cysteine sulfenic acid, Cysteinyl-selenocysteine, Deamidated asparagine, Deamidated glutamine, Dimethylated arginine, Diphthamide, Disulfide bond, GPI-anchor amidated alanine, GPI-anchor amidated asparagine, GPI-anchor amidated aspartate, GPI-anchor amidated cysteine, GPI-anchor amidated glycine, GPI-anchor amidated serine, Glutamic acid 1-amide, Glutamine amide, Glycine amide, Glycyl adenylate, Glycyl lysine isopeptide, Hydroxyproline, Hydroxyproline, Hypusine, Isoglutamyl cysteine thioester, Isoglutamyl lysine isopeptide, Isoleucine amide, Leucine amide, Leucine methyl ester, Lysine amide, Lysine tyrosylquinone, Methionine amide, N,N,N-trimethylalanine, N-acetylalanine, N-acetylaspartate, N-acetylcysteine, N-acetylglutamate, N-acetylglycine, N-acetylmethionine, N-acetylproline, N-acetylserine, N-acetylthreonine, N-acetylvaline, N-myristoyl glycine, N-palmitoyl cysteine, N-palmitoyl glycine, N-pyruvate 2-iminyl-valine, N4,N4-dimethylasparagine, N6,N6,N6-trimethyllysine, N6,N6-dimethyllysine, N6-(pyridoxal phosphate)lysine, N6-(retinylidene)lysine, N6-1-carboxyethyl lysine, N6-acetyllysine, N6-biotinyllysine, N6-carboxylysine, N6-lipoyllysine, N6-methylated lysine, N6-methyllysine, N6-myristoyl lysine, Nitrated tyrosine, O-(pantetheine 4'-phosphoryl)serine, O-AMP-threonine, O-AMP-tyrosine, O-acetylserine, O-acetylthreonine, O-decanoyl serine, O-palmitoyl serine, Omega-N-methylarginine, Omega-N-methylated arginine, Omega-hydroxyceramide glutamate ester, Phenylalanine amide, Phosphatidylethanolamine amidated glycine, Phosphohistidine, Phosphoserine, Phosphothreonine, Phosphotyrosine, PolyADP-ribosyl glutamic acid, Proline amide, Pyrrolidone carboxylic acid, Pyruvic acid, S-(dipyrrolylmethanemethyl)cysteine, S-8alpha-FAD cysteine, S-Lysyl-methionine sulfilimine, S-cysteinyl cysteine, S-farnesyl cysteine, S-geranylgeranyl cysteine, S-glutathionyl cysteine, S-methylcysteine, S-nitrosocysteine, S-palmitoyl cysteine, S-stearoyl cysteine, Sulfoserine, Sulfotyrosine, Symmetric dimethylarginine, Tele-8alpha-FAD histidine, Tele-methylhistidine, Thyroxine, Triiodothyronine, Tyrosine amide, Valine amide

From genome to proteome

genome

~ 20’000 protein coding-genes

proteome

~ 5'000'000 different proteins

alternative splicing of mRNA

2-5 fold increase

post-translational modifications of proteins (PTMs)

50-100 fold increase

transcriptome

~ 50 to 100’000 transcripts (mRNAs)

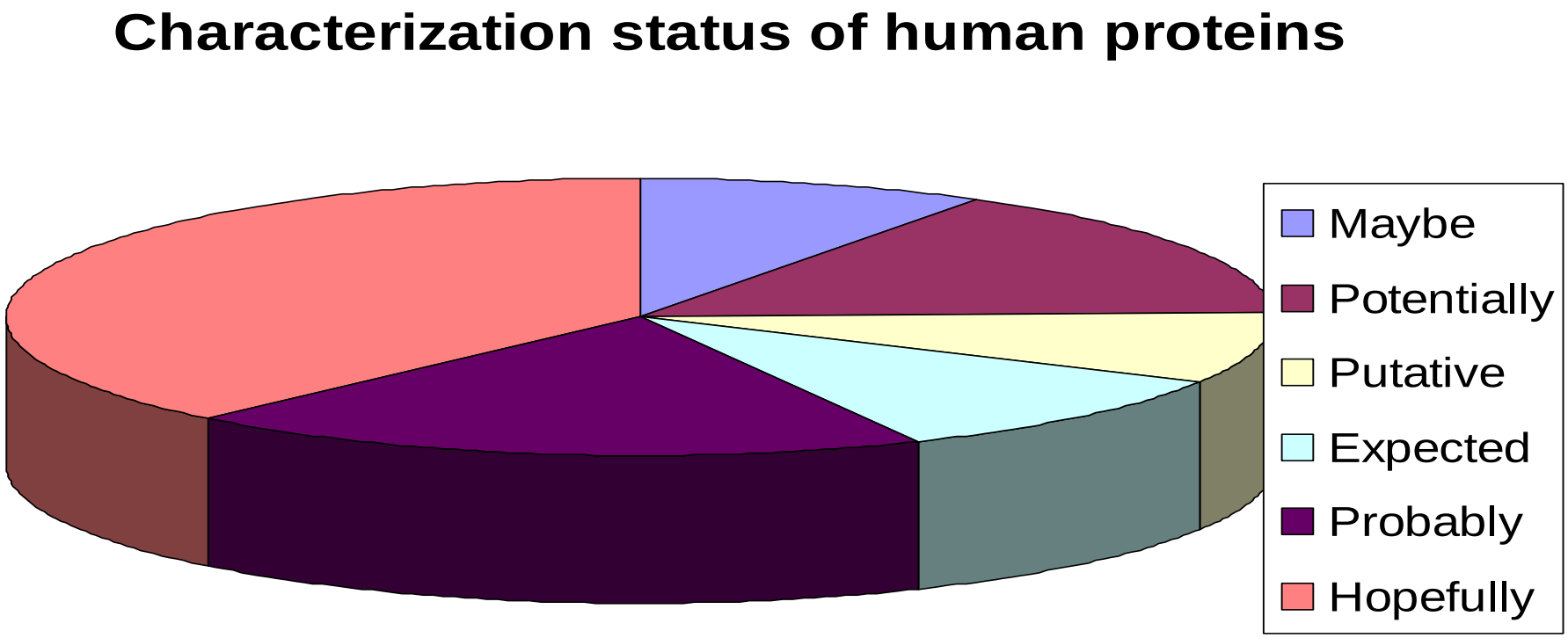
Protein complexity

Breakdown in term of Protein Evidence (PE) of human proteins

- 1: Evidence at protein level 13181 (65.0%)
- 2: Evidence at transcript level 6256 (30.8%)
- 3: Inferred from homology 213 (1.0%)
- 4: Predicted 103 (0.5%)
- 5: Uncertain 553 (2.7%)

But even for the 65% where there is evidence, at protein level, of the protein existence, there is still a lots to be done at the proteomic level (PTMs, interactions, subcellular location, tissue-specificity, etc).

In the framework of the annotation effort to produce a complete set of human entries, we were confronted by how little is known on the function of many human proteins….



Lots of human proteins with no or few clues on their functions

1. Similar to characterized proteins in distant organisms (bacteria, plants, yeast), but no validation in mammals;
2. Presence of domains that help predict a ‘general’ function but not a precise one (examples: hydrolase fold, GPCR);
3. Presence of domains or sequence features that help define some properties (examples: PDZ -> PPI, many TMs -> integral membrane protein);
4. “Orphan”. With no similarity to any characterized proteins but that can be conserved across a more or less wide taxonomic space.

About 5’000 human proteins are in one of the above categories

CALIPHO

Computer Analysis and Laboratory Investigation of Proteins of Human Origin

A new group of the University of Geneva and the Swiss Institute of Bioinformatics

Directed by Amos Bairoch and Lydie Lane

UNIVERSITÉ DE GENÈVE

SIB

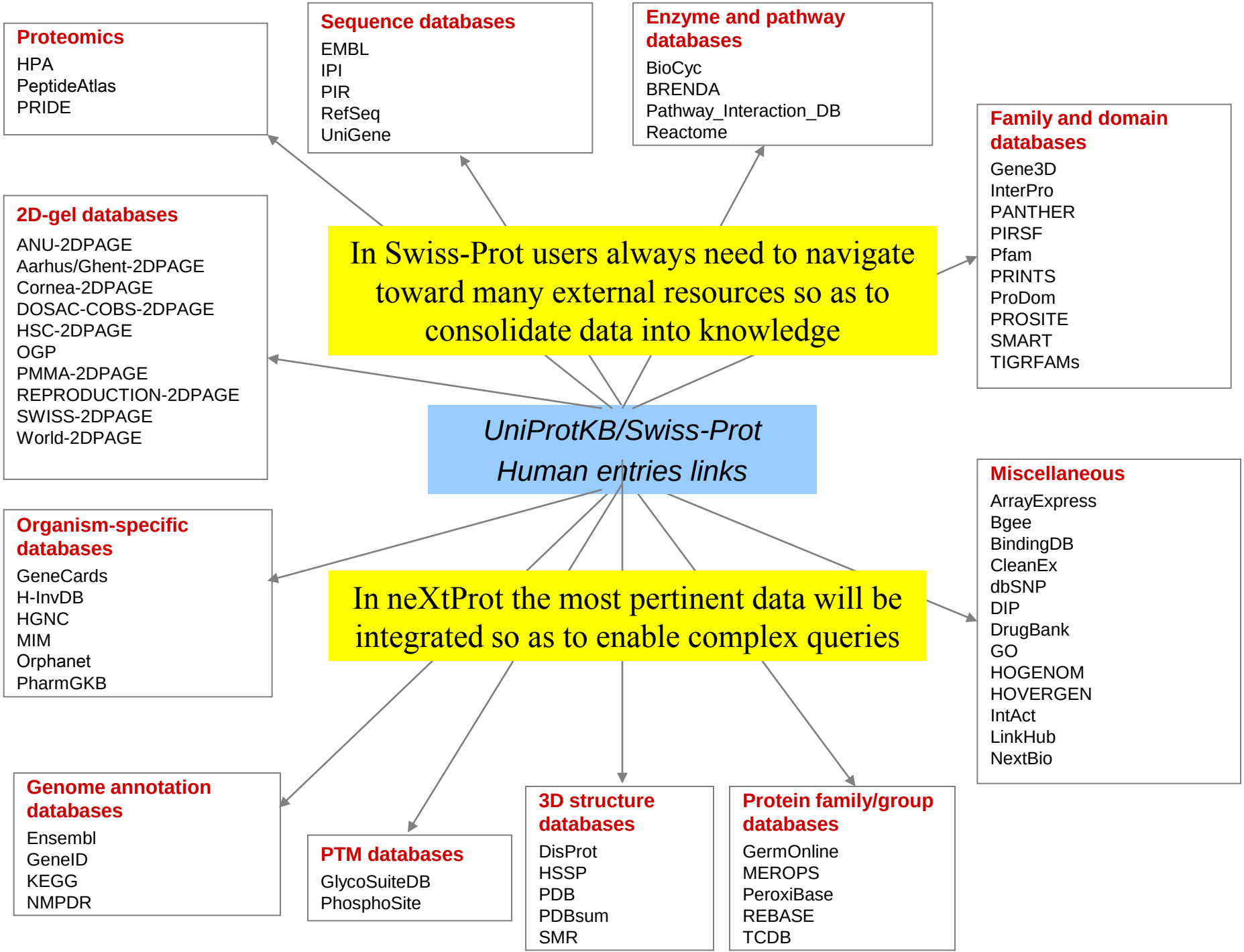
The 3 missions of CALIPHO

- Carry out laboratory experiments on selected sets of uncharacterized human proteins to discover their function;
- Develop **neXtProt**, an ambitious new knowledge resource centered around human proteins;
- Organize a collective effort that pools resources around the world with the goal of functionally characterize all human proteins.

neXtProt

- **What**: a comprehensive resource that complements SIB Swiss-Prot human protein annotation efforts. neXtProt is expected to become the central resource of human protein-centric information;
- **How**:
  - by mining, in the most appropriate way and with our stringent quality criteria, many external data resources.

    In this context we plan to add additional protein/protein interactions, proteomics data, pathway information, tissular and cellular expression from antibodies, variation data (such as SNP frequencies), siRNA screen data, microRNA targets, microarray expression data, phylogenetic profiling, etc;
  - by integrating experimental results from:
    - The new Geneva-based laboratory;
    - An extensive world-wide network of collaborators.

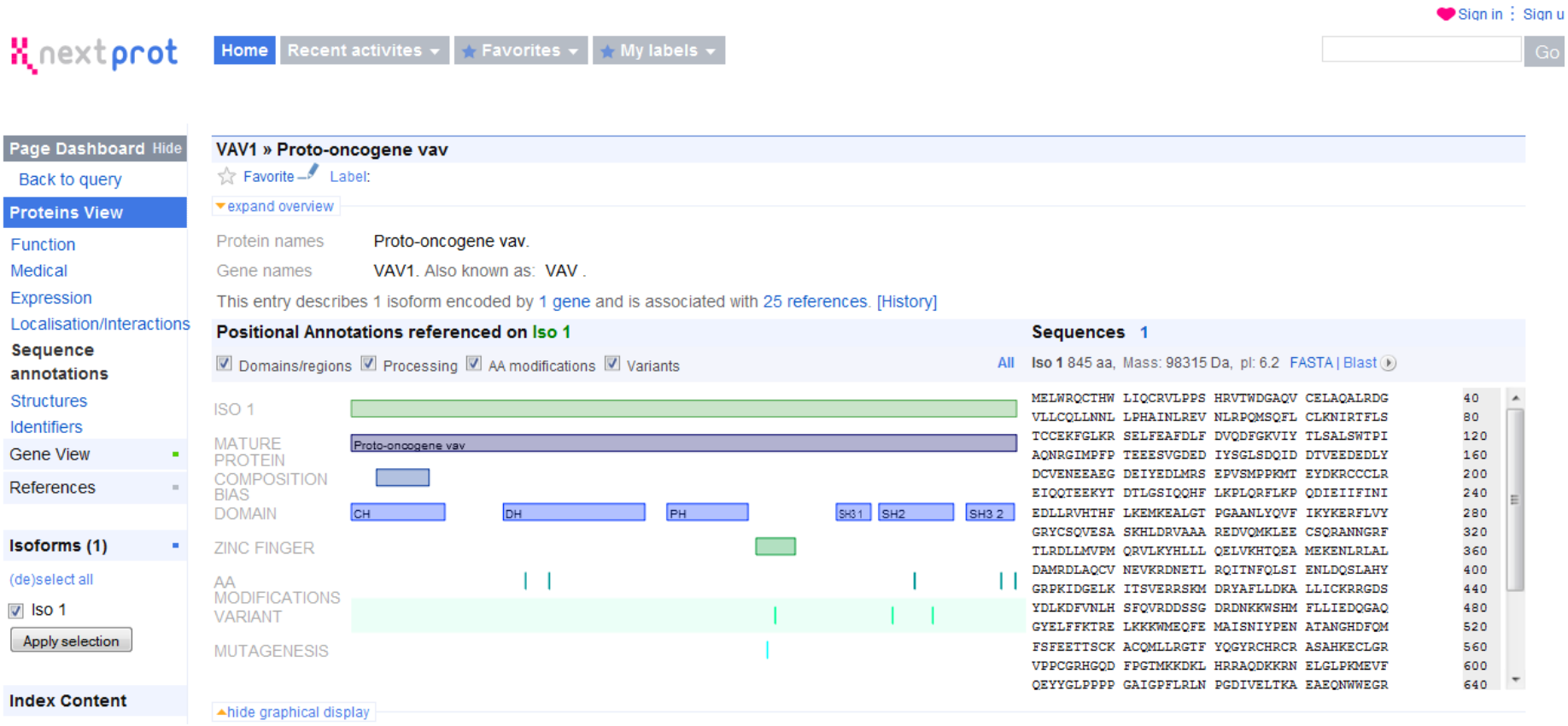

Proteomics
HPA
PeptideAtlas
PRIDE
Sequence databases
EMBL
IPI
PIR
RefSeq
UniGene
Enzyme and pathway databases
BioCyc
BRENDA
Pathway_Interaction_DB
Reactome
Family and domain databases
Gene3D
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs
2D-gel databases
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
Cornea-2DPAGE
DOSAC-COBS-2DPAGE
HSC-2DPAGE
OGP
PMMA-2DPAGE
REPRODUCTION-2DPAGE
SWISS-2DPAGE
World-2DPAGE
In Swiss-Prot users always need to navigate toward many external resources so as to consolidate data into knowledge
UniProtKB/Swiss-Prot
Human entries links
Miscellaneous
ArrayExpress
Bgee
BindingDB
CleanEx
dbSNP
DIP
DrugBank
GO
HOGENOM
HOVERGEN
IntAct
LinkHub
NextBio
Organism-specific databases
GeneCards
H-InvDB
HGNC
MIM
Orphanet
PharmGKB
In neXtProt the most pertinent data will be integrated so as to enable complex queries
Genome annotation databases
Ensembl
GeneID
KEGG
NMPDR
PTM databases
GlycoSuiteDB
PhosphoSite
3D structure databases
DisProt
HSSP
PDB
PDBsum
SMR
Protein family/group databases
GermOnline
MEROPS
PeroxiBase
REBASE
TCDB

What is not neXtProt?

- neXtProt is not Swiss-Prot “Plus”;
- Yes, neXtProt will contains a wealth of data not available in Swiss-Prot;
- But the real challenge is to build a real knowledge platform where our users can ask meaningful questions and hopefully obtain the answers that they seek!

When and what

- We will have a first version out in September 2010;
- It will already contain quite a number of innovative functionalities and some additional data sets: HPA, microarray data, SNPs frequencies, exons mapping, etc.

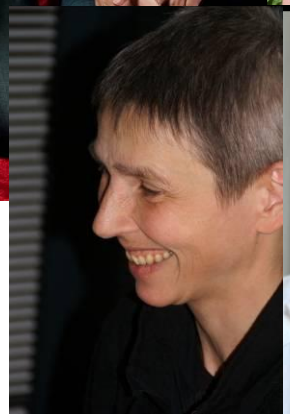


The future

- Our vision is to gradually build up neXtProt, not only by adding new data resources but:
  - By integrating state of the art data mining tools;
  - By integrating some forms of “social networking” functionalities allowing researches to share ideas and data;
  - By enabling the modeling of hypothesis inside the framework of the platform.

CALIPHO@UniGe_and_SIB

- **Laboratory:**
  - Franck Bontems, Aline Dousse, Camille Mary, Fabiana Tirone, Rachel Porcelli, Irene Rossito and Lisa Salleron
  - In close collaboration with: Richard Fish, Oliver Hartley
- **Biocurators:**
  - Guislaine Argoud-Puy, Isabelle Cusin, Paula Duek
- **Software developers:**
  - Olivier Evalet, Alain Gateau, Anne Gleizes, Catherine Zwahlen
  - Alexandre Masselot (GeneBio)
- **Research:**
  - Anais Mottaz
  - Anne-Lise Veuthey (Swiss-Prot), Marco Pagni (VitalIT)
- **Directed by:**
  - Amos Bairoch, Lydie Lane



The CALIPHO group