UNIVERSITÉ DE GENÈVE

SIB Swiss Institute of Bioinformatics

Bioinformatics for Human Proteomics: Current State and Future Status

Amos Bairoch

Tokyo, October 11, 2010

国立遺伝学研究所　国際シンポジウム

バイオデータベースの未来

Future perspectives of biological databases

ICEs

4th SIENA MEETING

FROM GENOME TO PROTEOME:
KNOWLEDGE ACQUISITION AND REPRESENTATION

Sept. 4-7, 2000, Siena, Italy

Exactly 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

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

[September 4, 2008]

A ‘complete’ set of annotated human proteins

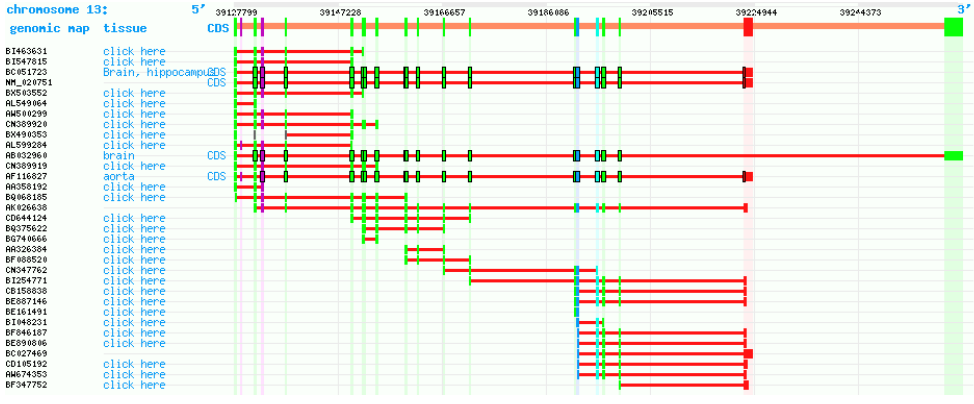
- In September 2008, we had annotated **20’330** human protein entries in UniProtKB/Swiss-Prot;
- They originate from about **20’400** protein-coding genes;
- Why ‘about’?
  - There are sets of genes that encode for identical proteins (example: 14 genes code for histone H4);
  - There are genes that 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.

Since...

- Since the beginning of 2009, we have added 130 «new» sequences, but we have «deleted» 171 proteins;
- Our gut feeling is that we are slowly but inexorably creeping toward slightly under 20’000 human-protein coding genes.

Alternative isoforms

- Produced by alternative splicing, promoter usage or initiation;
- Currently we have **14’700** additional isoforms in about **7’600** entries;
- This means that **38%** 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.



Sequence variants

- We have information concerning about **63’000** SAP (single amino-acid polymorphisms);
- **20’000** 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 sequence is artefactual. There is no such thing as an average genome;
- Some reported variants represent in fact the «majority» sequence;
- In humans, variability at level of proteins is not restricted to SAPs, one needs also to take into account:
  - Segregating pseudogenes (example: olfactory receptors);
  - Copy number variation;
  - Active retrotransposable elements (LINE-1, ERVs);
- The solution:to gradually move to what microbiologists have already embraced: a pangenome.

Post-translational modifications

- We have about **80’000** 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

alternative splicing of mRNA

2-5 fold increase

transcriptome

~ 50 to 100’000 transcripts (mRNAs)

proteome

~ 5'000'000 different proteins

post-translational modifications of proteins (PTMs)

50-100 fold increase

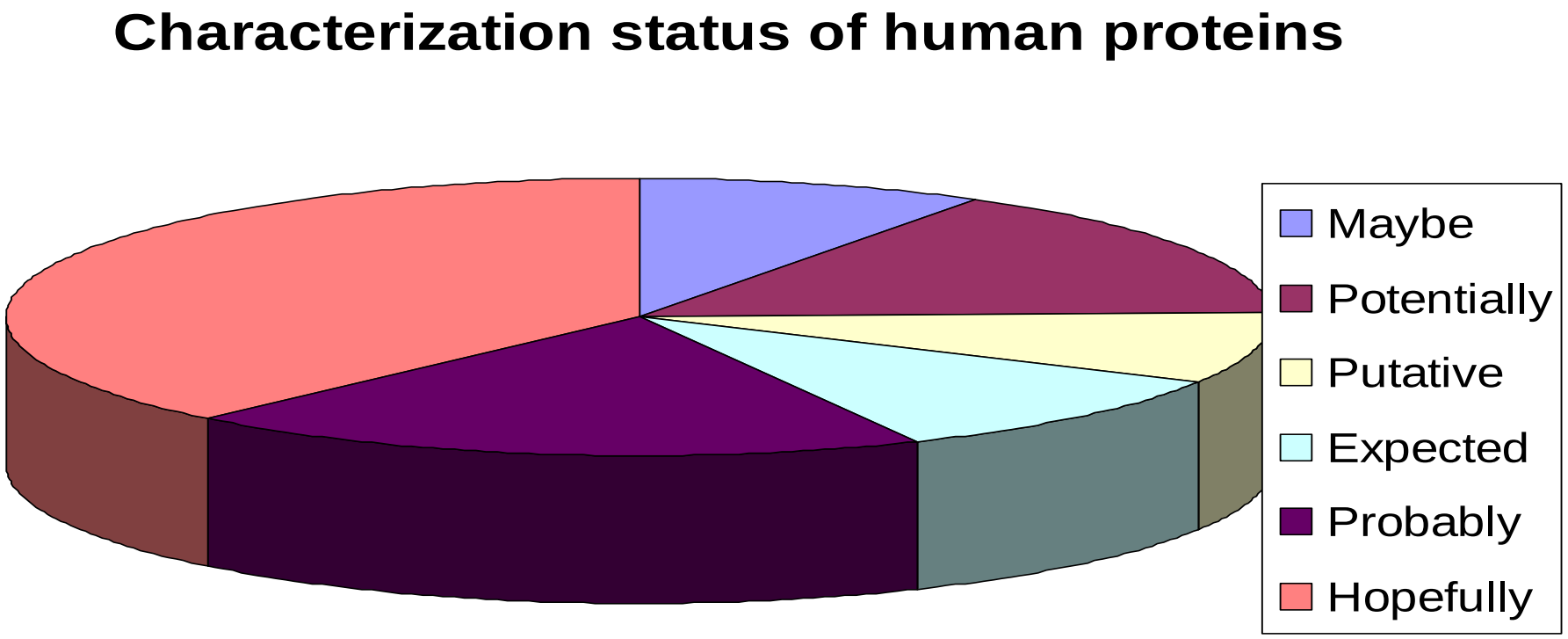
Protein complexity

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

1: Evidence at protein level	13340 (65.8%)
2: Evidence at transcript level	6018 (29.7%)
3: Inferred from homology	210 (1.0%)
4: Predicted	97 (0.5%)
5: Uncertain	600 (2.9%)

But even for the 66% 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….



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

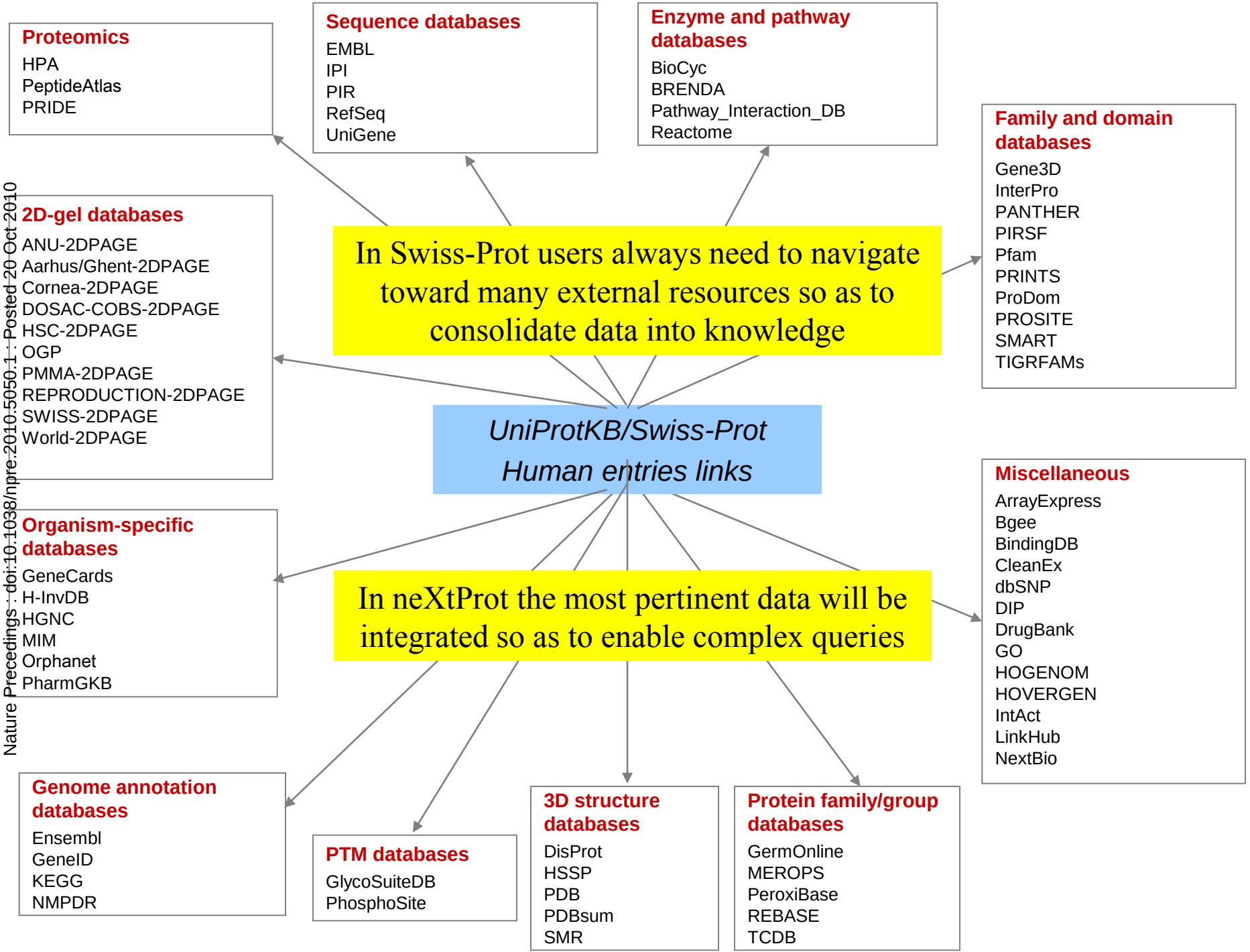
UNIVERSITÉ DE GENÈVE

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.

- **What**: a comprehensive resource that complements SIB/EBI 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 and protein/small molecules 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 an extensive network of collaborating laboratories.



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

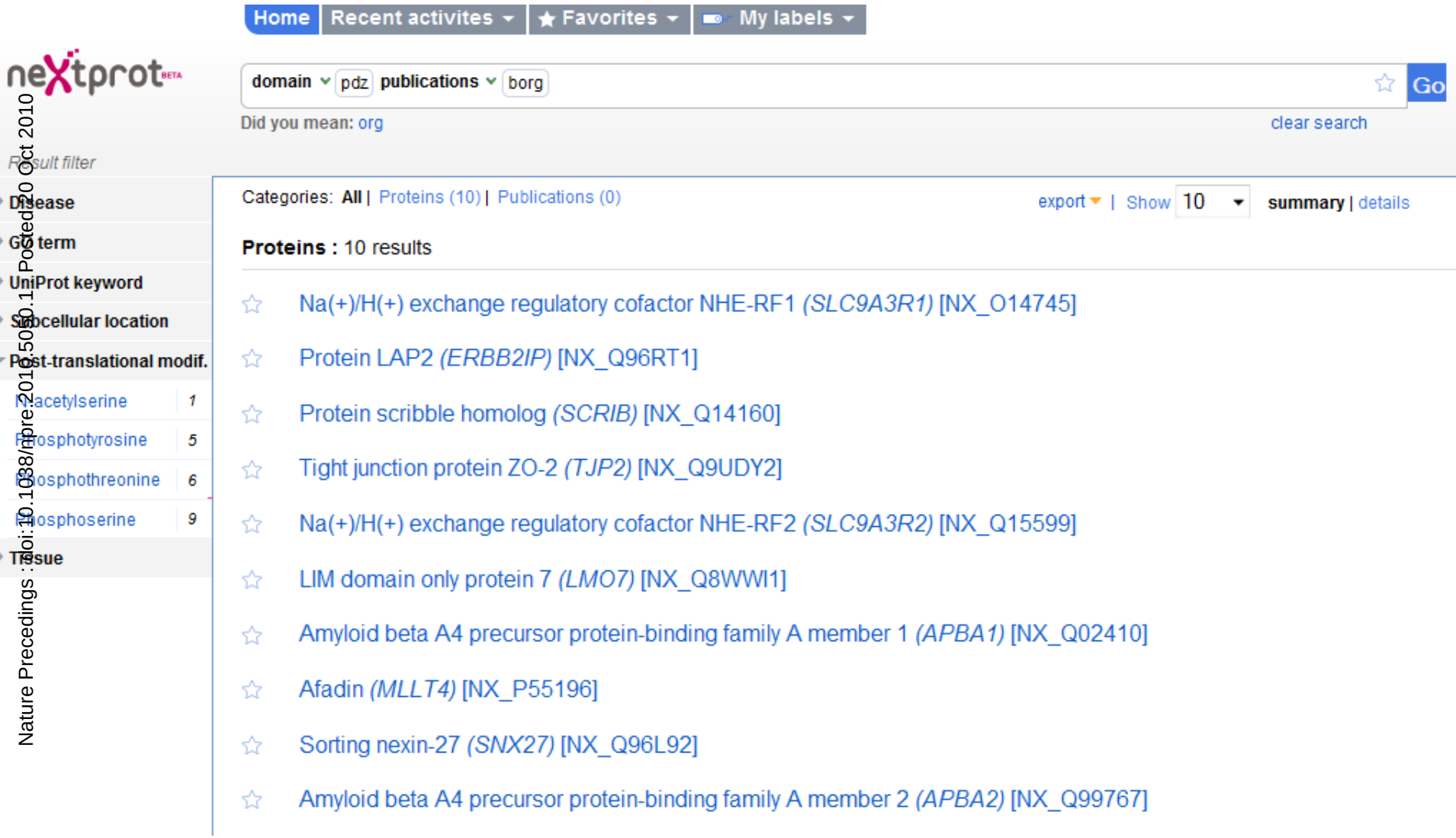
- We will have a first public version out in October 2010

In terms of data, it will contains:

- All of Swiss-Prot human data: sequences and annotations;
- Human Proteome Atlas organ and tissue expression information from antibodies;
- Metadata on mRNA expression from microarrays and ESTs from Bgee (analyzed from ArrayExpress and UniGene);
- Additional SNPs from dbSNP and Ensembl;
- Chromosomal location and exons mapping from Ensembl;
- Affymetrix and Illumina chip sets identifiers.

In terms of interface, it will offers:

- An intuitive query interface;
- Different specialized views (function, medical, expression, etc.);
- The possibility to tag and label proteins.



neXtprot BETA

everywhere | cadherin | Go

clear search

Result filter

- **Family name/EC number**
- **Disease**
  - Usher syndrome ty... 2
  - Focal segmental g... 1
  - Ectodermal dyspla... 1
  - Deafness autosoma... 1
  - Hypotrichosis and... 1
  - Hypotrichosis con... 1
  - Epidermolysis bul... 1
  - Palmoplantar kera... 1
  - Deafness autosoma... 1
  - Skin fragility-wo... 1
  - Usher syndrome ty... 1
  - Limb-girdle muscu... 1
- **GO term**
- **UniProt keyword**
- **Subcellular location**
- **Post-translational modif.**
- **Tissue**

Categories: **All** | Proteins (204) | Publications (490)

export | Show 10 | summary | **details**

Proteins : 10 of 204 results

☆ Cadherin-1 *(CDH1)* [NX_P12830]
E-Cad/CTF2 promotes non-amyloidogenic degradation of Abeta precursors. Has a strong inhibitory effect on APP C99 and C83 production.Cadherins are calcium-dependent cell adhesion proteins. [more]
Gene location: 16q22.1 Isoforms: 1 Variants: 35 PTMs: 6
3D structure: yes Proteomics: no Tissue expression: yes Mutagenesis: yes

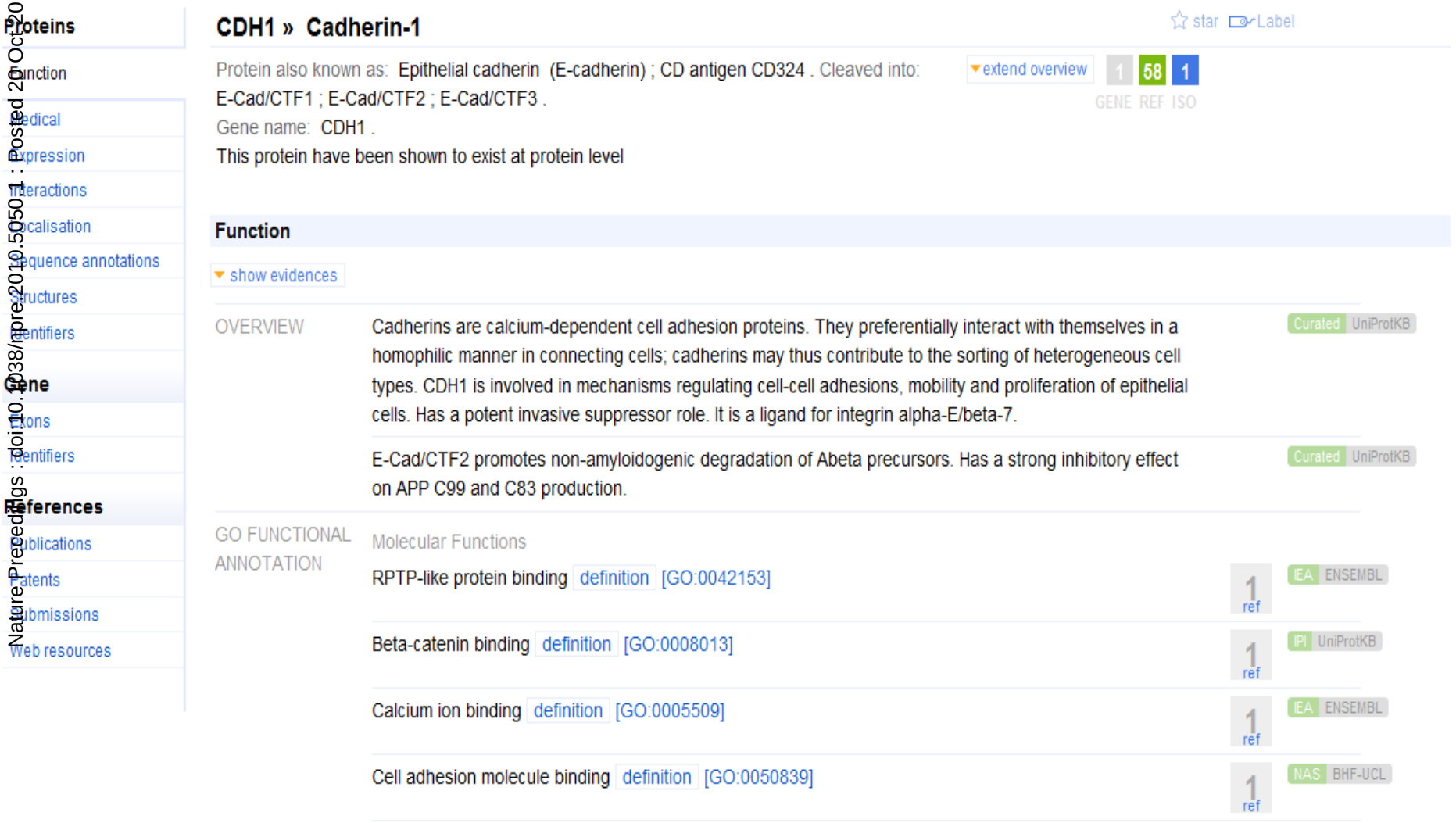
☆ Cadherin-3 *(CDH3)* [NX_P22223]
Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells; cadherins may thus contribute to the sorting of heterogeneous cell types.
Gene location: 16q22.1 Isoforms: 2 Variants: 15 PTMs: 3
3D structure: no Proteomics: no Tissue expression: yes Mutagenesis: no

☆ Cadherin-13 *(CDH13)* [NX_P55290]
Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells; cadherins may thus contribute to the sorting of heterogeneous cell types. May act as a negative regulator of neural cell growth.
Gene location: 16q23.3 Isoforms: 1 Variants: 2 PTMs: 9
3D structure: yes Proteomics: no Tissue expression: yes Mutagenesis: no

☆ Cadherin-23 *(CDH23)* [NX_Q9H251]
Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells. [more]
Gene location: 10q22.1 Isoforms: 6 Variants: 96 PTMs: 42
3D structure: yes Proteomics: no Tissue expression: yes Mutagenesis: no

☆ Cadherin-5 *(CDH5)* [NX_P33151]
Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells; cadherins may thus contribute to the sorting of heterogeneous cell types. [more]
Gene location: 16q21 Isoforms: 1 Variants: 5 PTMs: 8
3D structure: no Proteomics: no Tissue expression: yes Mutagenesis: no

A variety of views



VAV1 » Proto-oncogene vav

☆ star Label

Gene name: VAV1 .

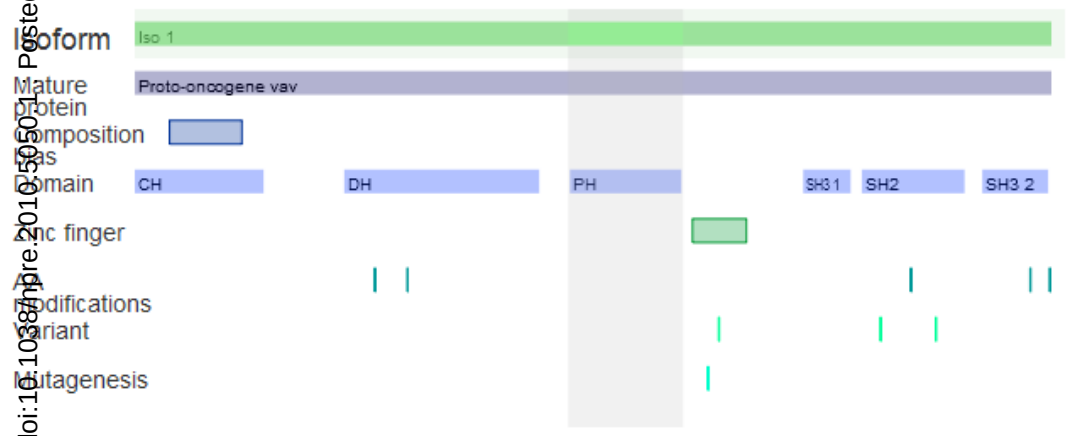
This protein have been shown to exist at protein level

extend overview 1 GENE 25 REF 1 ISO

Positional Annotations referenced on Iso 1

Isoform Iso 1 845 aa, Mass: 98314 Da, pI: 6.2

Domains/regions Processing AA modifications Variants All/None Actions: FASTA , Blast: full sequence on selection




```
241 EDLLRVHTHF LKEMKEALGT PGAANLYQVF IKYKERFLVY
281 GRYCSQVESA SKHLDRVAAA REDVQMKLEE CSQRANNGRF
321 TLRDLLMVPM QRVLKYHLLL QELVKHTQEA MEKENLRLAL
361 DAMRDLAQCV NEVKRDNETL RQITNFQLSI ENLDQSLAHY
401 GRPKIDGELK ITSVERRSKM DRYAFLLDKA LLICKRRGDS
441 YDLKDFVNLH SFQVRDDSSG DRDNKKWSHM FLLIEDQGAQ
481 GYELFFKTRE LKKKWMEQFE MAISNIYPEN ATANGHDFQM
521 FSFEETTSCK ACQMLLRGTF YQGYRCHRCR ASAHKECLGR
561 VPPCGRHGQD FPGTMKKDKL HRRAQDKKRN ELGLPKMEVF
601 QEYYGLPPPP GAIGPFLRLN PGDIVELTKA EAEQNWWEGR
641 NTSTNEIGWF PCNRVKPYVH GPPQDLSVHL WYAGPMERAG
681 AESILANRSD GTFLVRQRVK DAAEFAISIK YNVEVKHIKI
721 MTAEGLYRIT EKKAFRGLTE LVEFYQQNSL KDCFKSLDTT
761 LQFPFKEPEK RTISRPAVGS TKYFGTAKAR YDFCARDRSE
801 LSLKEGDIIK ILNKKGQQGW WRGEIYGRVG WFPANYVEED
841 YSEYC
```

Hide graphical display

Category	Names	Positions	Length	Description	Evidences	Also present in isoforms
	Domain	402 - 504	103	PH	1	
	Zinc finger	515 - 564	50	Phorbol-ester/DAG-type	1	
	Domain	617 - 660	44	SH3 1	1	
	Domain	671 - 765	95	SH2	1	
	Domain	782 - 842	61	SH3 2	1	
PROCESSING	Mature protein	1 - 845	845	Proto-oncogene vav	1	

Biophysicochemical properties

Kinetic parameters

K_M	663.5 uM for D-methylphenidate
K_M	43 mM for ethanol
K_M	116 uM for cocaine
K_M	775.7 uM for L-methylphenidate
K_M	106.6 uM for p-nitrophenyl acetate
V_{max}	493.9 nmol/min/mg enzyme with p-nitrophenyl acetate as substrate
V_{max}	177.2 pmol/min/mg enzyme with D-methylphenidate as substrate
V_{max}	1701.1 pmol/min/mg enzyme with L-methylphenidate as substrate

Dependence

pH	Optimum pH is 6.5.

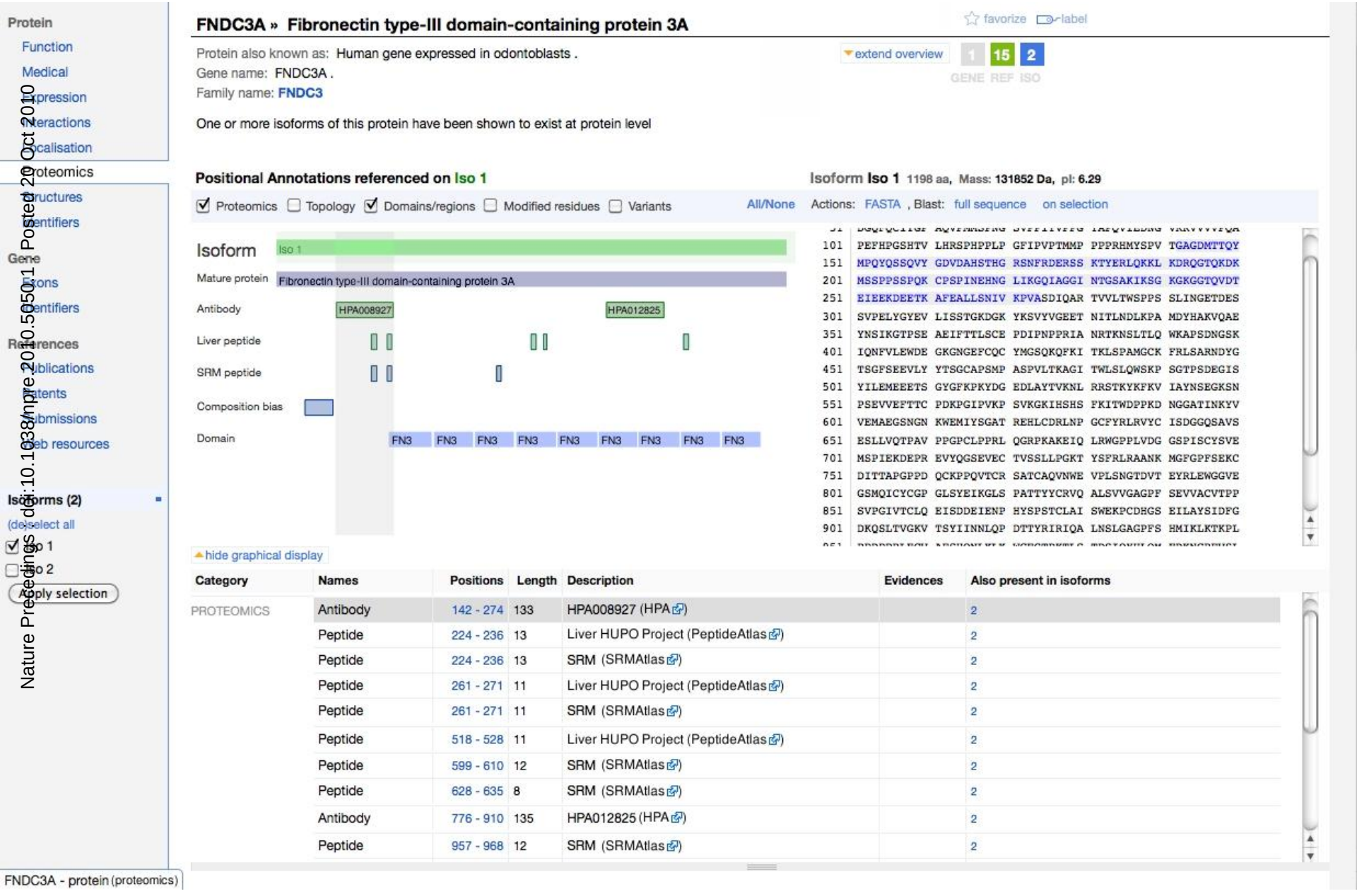
Tissue expression

Tissue	Expressed at mRNA level			Expressed at protein level	
	Chips		ESTs	Antibodies	
Alimentary system		29	-		18
|→Gastrointestinal tract		23	-		13
| |→Intestine		11	-		6
| | |→Intestinal mucosa		-	-		5
| | | |→Intestinal epithelium		-	-		5
| | | | |→Large intestine epithelium		-	-		3
| | | | | |→Colon epithelium		-	-		1

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.
- To work closely with HUPO HPP stakeholders to define what proteomics-derived data we will represent in neXtProt.

A prototype of a future proteomics view



You can already start test drive neXtProt in a few days

Just go to
beta.nextprot.org
and sign up

CALIPHO@UniGe_and_SIB

- **neXtProt content:**
  - Coordinator: Pascale Gaudet
  - Biocurators: Guislaine Argoud-Puy, Isabelle Cusin, Paula Duek
- **neXtProt software developers:**
  - Olivier Evalet, Alain Gateau, Anne Gleizes, Catherine Zwahlen and Alexandre Masselot (GeneBio)
- **Bioinformatics research:**
  - Anais Mottaz, Anne-Lise Veuthey (Swiss-Prot), Marco Pagni (VitalIT)
- **Laboratory research:**
  - Franck Bontems, Marjorie Desmurs, Camille Mary, Fabiana Tirone, Rachel Porcelli, Irene Rossito and Lisa Salleron
- **Directed by:**
  - Amos Bairoch, Lydie Lane and Nasri Nahas

The CALIPHO group