

Integrating Public and Private Data: Lessons Learned from Unison

<http://unison-db.org/>

Online access, download, documentation, references.

Reece Hart
Genentech, Inc.

Molecular Medicine Tri-Conference
February 26, 2009
San Francisco, CA

Genentech
IN BUSINESS FOR LIFE

Updates available at <http://harts.net/reece/pubs/>

A Bestiary of Life Sciences Data Types

Genomics

assemblies, transcripts,
probes, trans. factors,
expression, SNPs,
haplotypes

Proteomics

sequences, domains, PTMs,
localization, structure,
orthology, predictions

Chemistry

compounds, HCS, HTS,
properties

Networks

interactions, pathways

Communications

literature, patents, and
presentations

LIMS

animal records, protocols
request systems,
personnel, samples

Clinical

assays, protocols,
patient records,
samples

Annotation

GO, taxonomy, SCOP,
disease, OMIM

Types of Integration

➤ **Semantic Integration**

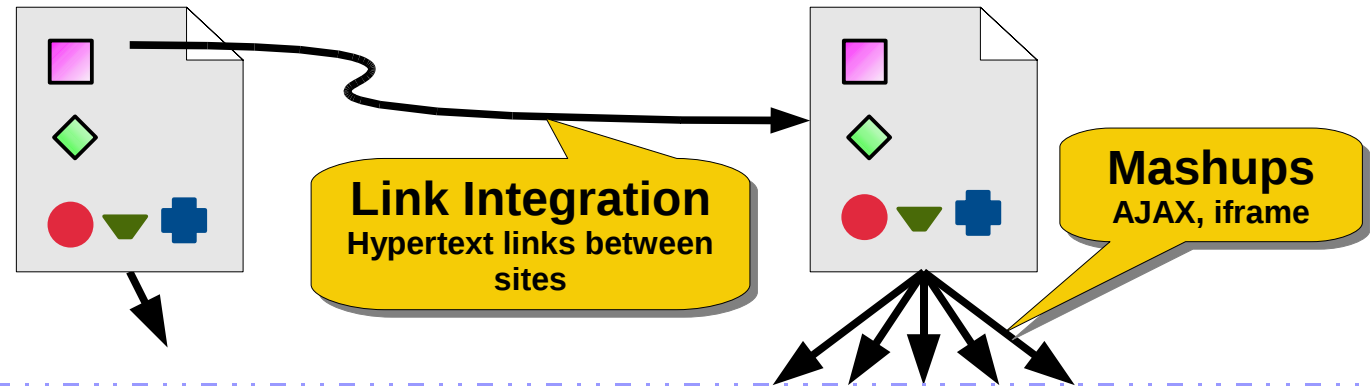
- Integrates fundamentally distinct data types.
- Improves contextual understanding of data.

➤ **Source Aggregation**

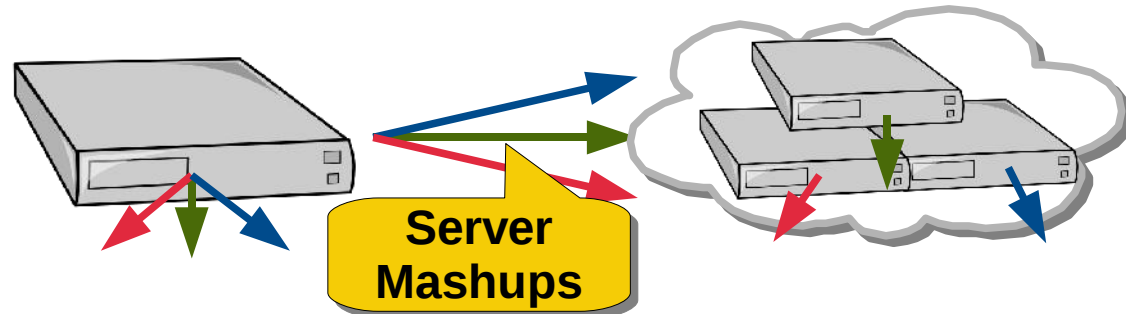
- Aggregates data of the same type from multiple sources. *e.g.*, in-house sequences with external ones.
- Ensures completeness of data.

A Survey of Integration Methods

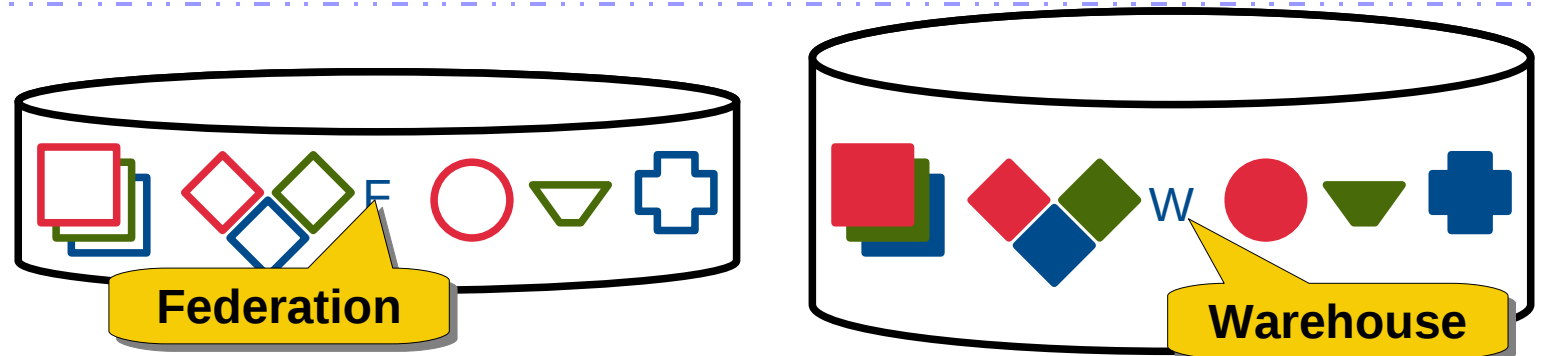
Presentation



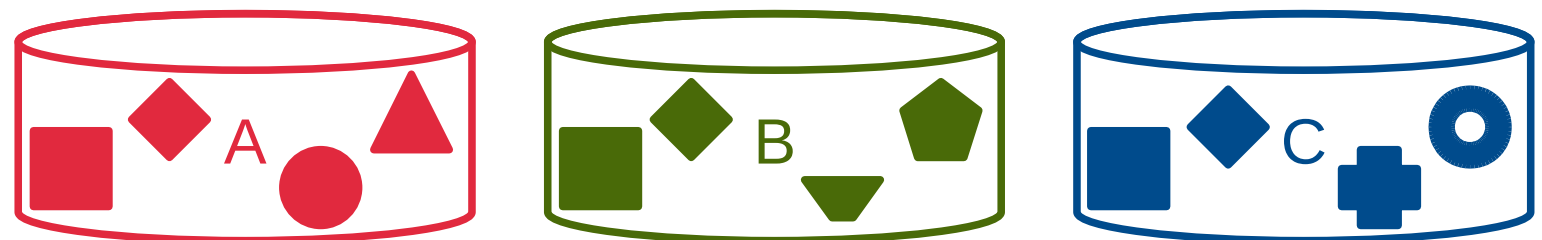
Middle Tier



Database Integration
(Federation / Warehouse)



Source Databases
or Files



Why is Integration Difficult?

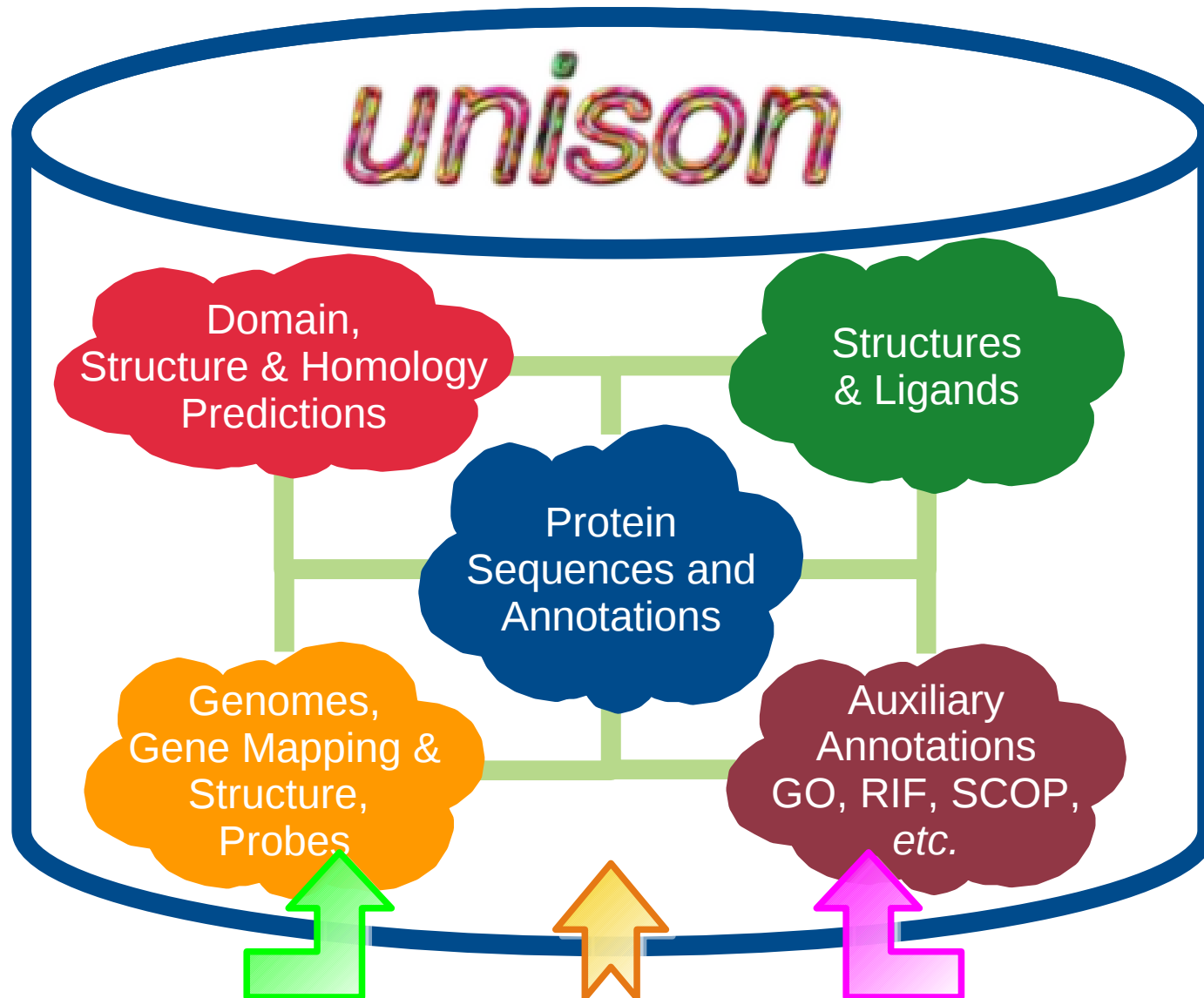
- **Establishing semantic equivalences and relationships are difficult.**
- **Source databases are updated often.**
 - Volume and frequency of updates are challenging.
- **Source databases have dynamic structure.**

Benefit Lessons

- ***Integrate to enable reasoning based on a corpus of data of multiple types and/or from multiple origins.***
 - To analyze biological data in broad context.
 - To generate hypotheses by data mining.
 - To enable business decisions based on a holistic view of decision criteria.

- **Ancillary benefits:**
 - Data preparation is hard. Centralization means that questions get asked and asked efficiently.
 - Integrated data provides a consistent foundation on which others can build.
 - Integration improves currency.

Unison in a Nutshell



Sequences and Annotations

UniProt, IPI, Ensembl, RefSeq, PDB, PHANTOM, HUGE, ROUGE, MGC, Derwent, pataa, nr, etc.
>13M seqs, >17k species, 69 origins

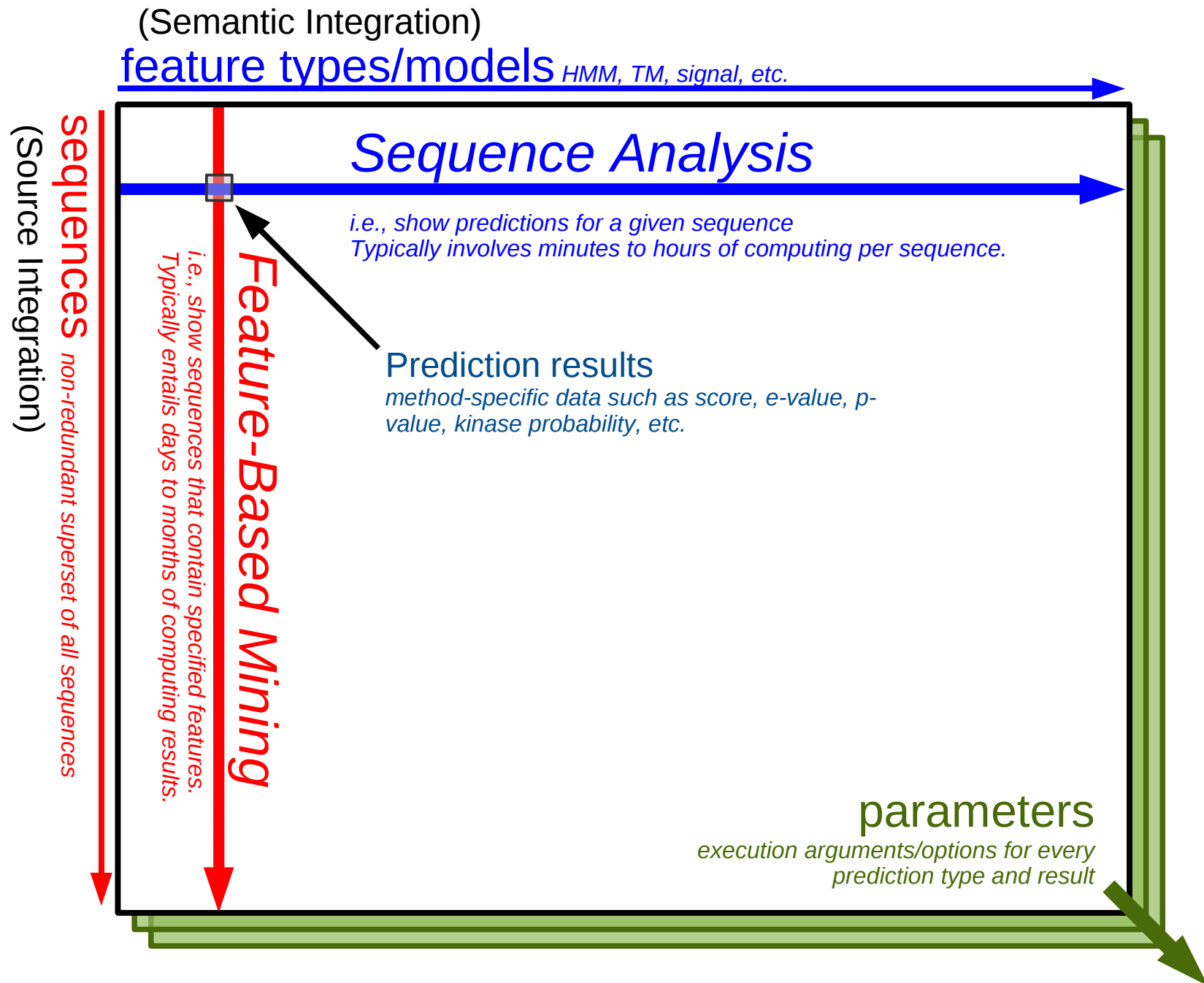
Auxiliary Data

HomoloGene, Gene Ontology, taxonomy, PDB, HUGO, SCOP, etc.

Precomputed predictions

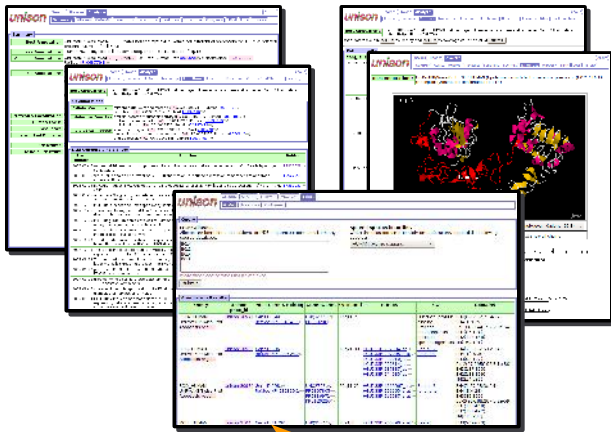
Domains, homology, structure, TMs, localization, signals, disorder, etc.
>200M predictions, 23 types, ~6 CPU-years

Analysis and Data Mining Have Distinct Needs.

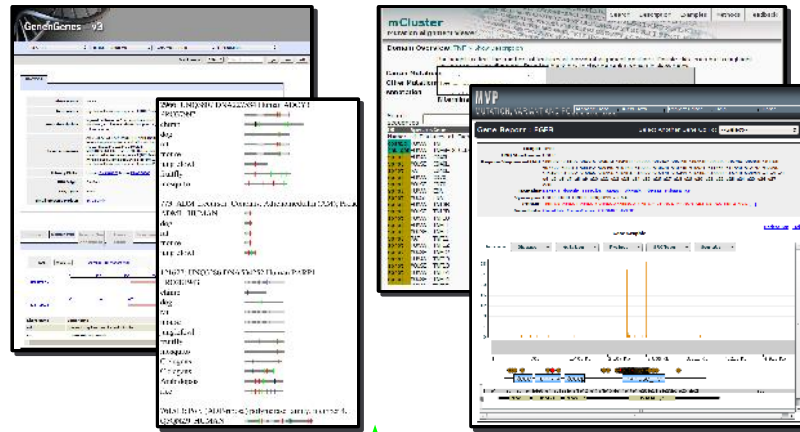


Unison has many applications.

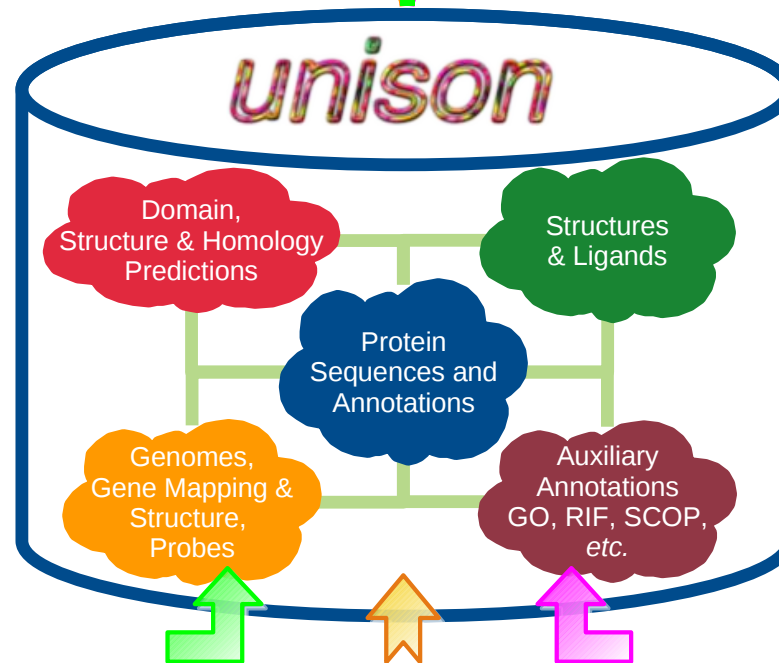
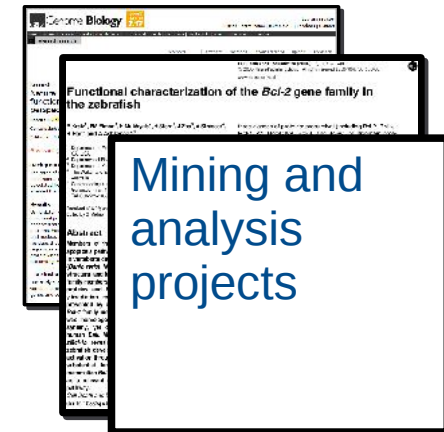
Unison Web Tools



Other In-House Tools



Ad Hoc Mining

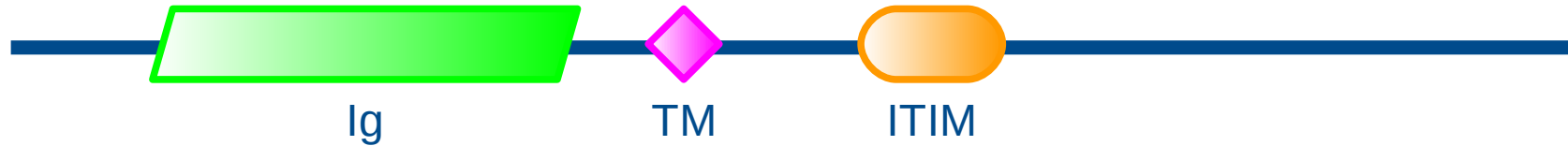


Sequences and Annotations
 UniProt, IPI, Ensembl, RefSeq, PDB
 STRING, PHANTOM, HUGE, ROUGE,
 MGC, Derwent, pataa, nr, etc.
 >13M seqs, >17k species, 69 origins

Auxiliary Data
 HomoloGene, Gene
 Ontology, taxonomy,
 PDB, HUGO, SCOP,
 etc.

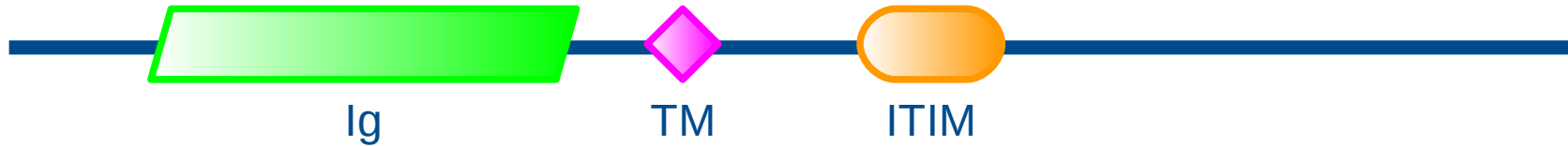
Precomputed predictions
 Domains, homology, structure, TMs,
 localization, signals, disorder, etc.
 >200M predictions, 23 types,
 ~6 CPU-years

Mining for ITIMs the old way.



- **Collect sequences.**
- **Prune redundant sequences. (How?!)**
- **For each unique sequence, predict**
 - Immunoglobulin domains.
 - Transmembrane domains.
 - ITIM domains.
- **Write a program that filters predictions.**
- **Summarize hits with external data.**
- **Do it again when source data are updated.**

Mining for ITIMs the Unison way.

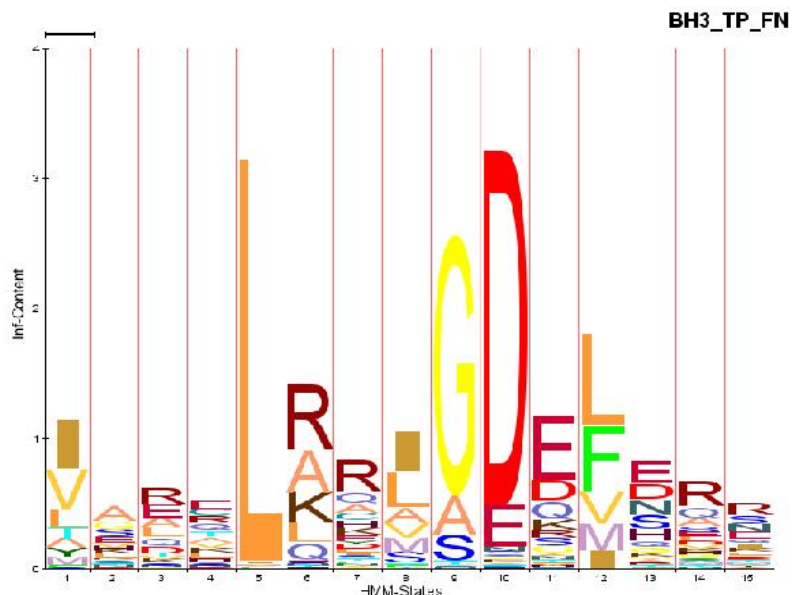


```

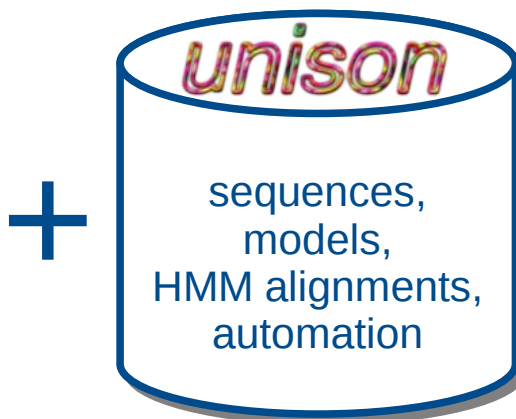
SELECT IG.pseq_id,
       IG.start as ig_start,IG.stop as ig_stop,IG.score,IG.eval,
       TM.start as tm_start,TM.stop as tm_stop,
       ITIM.start as itim_start,ITIM.stop as itim_stop
FROM   pahmm_current_pfam_v IG
JOIN   pftmhmm_tms_v TM ON IG.pseq_id=TM.pseq_id  AND IG.stop<TM.start
JOIN   pfregex_v ITIM  ON TM.pseq_id=ITIM.pseq_id AND TM.stop<ITIM.start
WHERE  IG.name='ig' AND IG.eval<1e-2
       AND ITIM.acc='MOD_TYR_ITIM';
    
```

<u>pseq_id</u>	<u>Ig</u> <u>start</u>	<u>Ig</u> <u>stop</u>	<u>score</u>	<u>eval</u>	<u>TM</u> <u>start</u>	<u>Tm</u> <u>stop</u>	<u>ITIM</u> <u>start</u>	<u>ITIM</u> <u>stop</u>	<u>best annotation</u>
234	262	316	30	7.40E-06	440	462	518	523	UniProtKB/Swiss-Prot:SIGL5_HUMAN (Rec
254	158	213	36	1.90E-07	284	306	386	391	UniProtKB/Swiss-Prot:VSIG4_HUMAN (Rec
544	157	215	24	6.60E-04	348	370	431	436	UniProtKB/Swiss-Prot:SIGL9_HUMAN (Rec
797	254	312	40	7.60E-09	1099	1121	1361	1366	UniProtKB/Swiss-Prot:DCC_HUMAN (RecN
1113	42	102	30	1.20E-05	243	265	300	305	UniProtKB/Swiss-Prot:KI2L2_HUMAN (RecM
1114	42	102	30	6.50E-06	243	265	330	335	UniProtKB/Swiss-Prot:KI2L1_HUMAN (RecM
1115	42	102	31	4.20E-06	243	265	301	306	UniProtKB/Swiss-Prot:KI2L3_HUMAN (RecM
1116	42	97	30	1.10E-05	339	361	396	401	UniProtKB/TrEMBL:Q95368_HUMAN (SubN
1134	340	388	26	1.40E-04	603	625	688	693	UniProtKB/Swiss-Prot:PECA1_HUMAN (Re

Data Integration Led to Bcl-2 Discoveries.

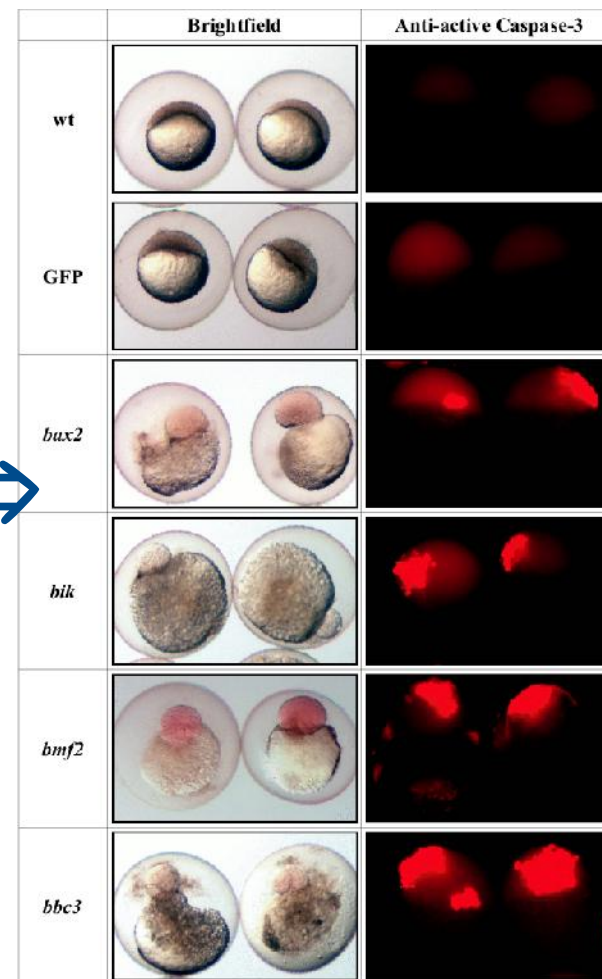


Custom model building



Zfish Protein	Source Database and Accession	Human Protein	E-value	Score	% Ide	% Coverage
Bax	RefSeq:NP_571637	BAX	2.00E-47	189	51	98
Bax2	E35:ENSDARP00000040899		1.00E-14	81	33	51
Bik	UP:Q5RGV6_BRARE	BIK	1.41E+04	20	47	12
Bmf	RefSeq:NP_001038689	BMF	1.00E-05	50	32	91
Bmf2	E35:FGENESH00000082230		1.10E-02	42	41	42
BBC3	E35:FGENESH00000078270	PUMA	2.10E+01	30	25	49

4 novel Bcl-2 proteins in zebrafish



Unison Web Tools

unison Search Browse Analyze About

Summary Aliases Patents Homologs Functions Features Structure Prospect HMM Loci History

Summary

Best Annotation UniProtKB/Swiss-Prot:MCL1_HUMAN (Induced myeloid leukemia cell differentiation protein Mcl-1 (Bcl-2- related protein EAT/mcl1) (mcl1/EAT).)

Entrez Annotations Human MCL1; myeloid cell leukemia sequence 1 (BCL2-related) (1q21)

Common Annotations UniProtKB/Swiss-Prot:[MCL1_HUMAN](#) ⇨, UniProtKB/Swiss-Prot:[Q07820](#) ⇨, GenenGenes:[PRO37003](#) ⇨, RefSeq:[NP_068779.1](#) ⇨
See the [Aliases](#) tab for all aliases.

Go Annotations Component: mitochondrial outer membrane
Component: cytoplasm
Function: protein heterodimerization activity
Function: protein channel activity
Function: protein binding
Process: apoptotic program
Process: multicellular organismal development
Process: anti-apoptosis
See the [Functions](#) tab for evidence, PubMed references, and NCBI GeneRIFs.

Protcomp Localization Plasma membrane by sequence similarity

Human Locus 1q21.2

Sequence 350 AA (MFGLKRNAVIGLNLY...AGVAGVGAGLAYLIR) [download FASTA](#)

Predicted Domains Domain Digest: BH3(209-223), Bcl-2(213-312;170;7.3e-48), BH1(253-272), BH2(305-316), TM(327-349)
Phosphorylation sites (pos;probability): pSer(25;0.914)

Structures [1wsx](#) ⇨, [2jm6](#) ⇨, [2nla](#) ⇨, [2pqq](#) ⇨

Domain Structure

Unison:2104; 350 AA; UniProtKB/Swiss-Prot:MCL1_HUMAN

SignalP
NN (0.27)
HMM (0.10)

Unison Web Tools

The screenshot shows the Unison web tool interface. At the top, there is a navigation bar with buttons for 'Search', 'Browse', 'Analyze', and 'About'. Below this is a secondary navigation bar with buttons for 'Summary', 'Aliases', 'Patents', 'Homologs', 'Functions', 'Features', 'Structure', 'Prospect', 'HMM', 'Loci', and 'History'. The main content area is titled 'Best Annotation?' and displays the UniProtKB/Swiss-Prot entry for MCL1_HUMAN. Below this, there are sections for 'Go Annotations', 'Cellular Component', 'Molecular Function', and 'Biological Process', each listing associated Gene Ontology (GO) terms and PubMed IDs. At the bottom, there is a section for 'NCBI GeneRIFs & References' which contains a table of references.

Best Annotation ? UniProtKB/Swiss-Prot:MCL1_HUMAN (Induced myeloid leukemia cell differentiation protein Mcl-1 (Bcl-2- related protein EAT/mcl1) (mcl1/EAT).)

Go Annotations

Cellular Component mitochondrial outer membrane (TAS; GO:0005741; PubMed:[10837489](#)↗)
cytoplasm (TAS; GO:0005737; PubMed:[10837489](#)↗)

Molecular Function protein heterodimerization activity (IPI; GO:0046982; PubMed:[10837489](#)↗)
protein channel activity (TAS; GO:0015266; PubMed:[10837489](#)↗)
protein binding (IDA; GO:0005515; PubMed:[10837489](#)↗)

Biological Process apoptotic program (TAS; GO:0008632; PubMed:[7682708](#)↗)
multicellular organismal development (TAS; GO:0007275; PubMed:[8790944](#)↗)
anti-apoptosis (TAS; GO:0006916; PubMed:[10837489](#)↗)

NCBI GeneRIFs & References

Last Update	Function	PubMed
2008-03-15	Akt and Mcl-1 are major components of a survival pathway that can be activated in CLL B cells by antigen stimulation.	17928528 ↗
2008-03-15	While TNFalpha had no effect on MCL-1 transcription, it induced expression of another antiapoptotic molecule, BFL-1.	17942758 ↗
2008-01-05	BCR/ABL induces SPK1 expression and increases its cellular activity, leading to upregulation of Mcl-1 in CML cells.	17599053 ↗
2008-01-05	a caspase-9 signaling cascade induces feedback disruption of the mitochondrion through cleavage of anti-apoptotic Bcl-2, Bcl-xL, and Mcl-1	17893147 ↗
2007-12-22	MCL1 determines the Bax dependency of Nbk/Bik-induced apoptosis.	18025305 ↗
2007-12-15	These results suggest that the N terminus of MCL-1 plays a major regulatory role, regulating coordinately the mitochondrial (anti-apoptotic) and nuclear (anti-proliferative) functions of MCL-1.	17823113 ↗
2007-12-08	Mcl-1 degradation primes the cell for Bim and Bax activation and anoikis, which can be blocked by oncogenic signaling in metastatic cells.	18006817 ↗

Unison Web Tools

unison

Search Browse **Analyze**

About

Summary Aliases **Patents** Homologs Functions Features Structure Prospect HMM Loci History

Best Annotation ? UniProtKB/Swiss-Prot:MCL1_HUMAN (Induced myeloid leukemia cell differentiation protein Mcl-1 (Bcl-2- related protein EAT/mcl1) (mcl1/EAT).)

show patents within % identity and % coverage of Unison:2104

Patent Results

pseq_id	len	%IDE	%COV	alias	species	date	authority	description
1829475	350	100	100	Geneseq:AAR68814	Homo sapiens	2003-03-25	UNIV JOHNS HOPKINS SCHOOL MED.	W09429330-A1; New myeloid cell leukaemia associated gene mcl-1 - used to develop prodsfor detection and treatment of cell proliferative disorders, partic. myeloid cell leukaemia. [DT: 25-MAR-2003 (revised); 15-JUL-1995 (first entry)] [PA: (UYJO) UNIV JOHNS HOPKINS SCHOOL MED.] [PI: Craig RW;] [OS: Homo sapiens]
122462	350	100	100	Geneseq:ADE25741	Homo sapiens	2004-01-29	INCYTE GENOMICS INC.	US2003194721-A1; Combination containing several polynucleotide that are differentially expressed in foam cells and complements of the polynucleotides, useful for diagnosing cardiovascular disease or atherosclerosis [DT: 29-JAN-2004 (first entry);] [PA: (INCY-) INCYTE GENOMICS INC.] [PI: Mikita T, Shiffman D, Porter JG, Kaser MR;] [OS: Homo sapiens]
2104	350	100	100	Geneseq:ADN03670	Homo sapiens	2004-07-01	GENENTECH INC.	W02004028479-A2; New PRO nucleic acid or polypeptide, useful for preparing a pharmaceutical composition for diagnosing or treating psoriasis in a mammal [DT: 01-JUL-2004 (first entry);] [PA: (GETH) GENENTECH INC.] [PI: Bodary S, Clark H, Jackman J, Schoenfeld J, Williams PM, Wood W;Wu TD;] [OS: Homo sapiens]

Unison Web Tools

unison

SEARCH

BROWSE

ANALYZE

ABOUT

SUMMARY

ALIANES

PATENTS

HOMOLOGS

FUNCTIONS

FEATURES

STRUCTURE

PROSPECT

HMM

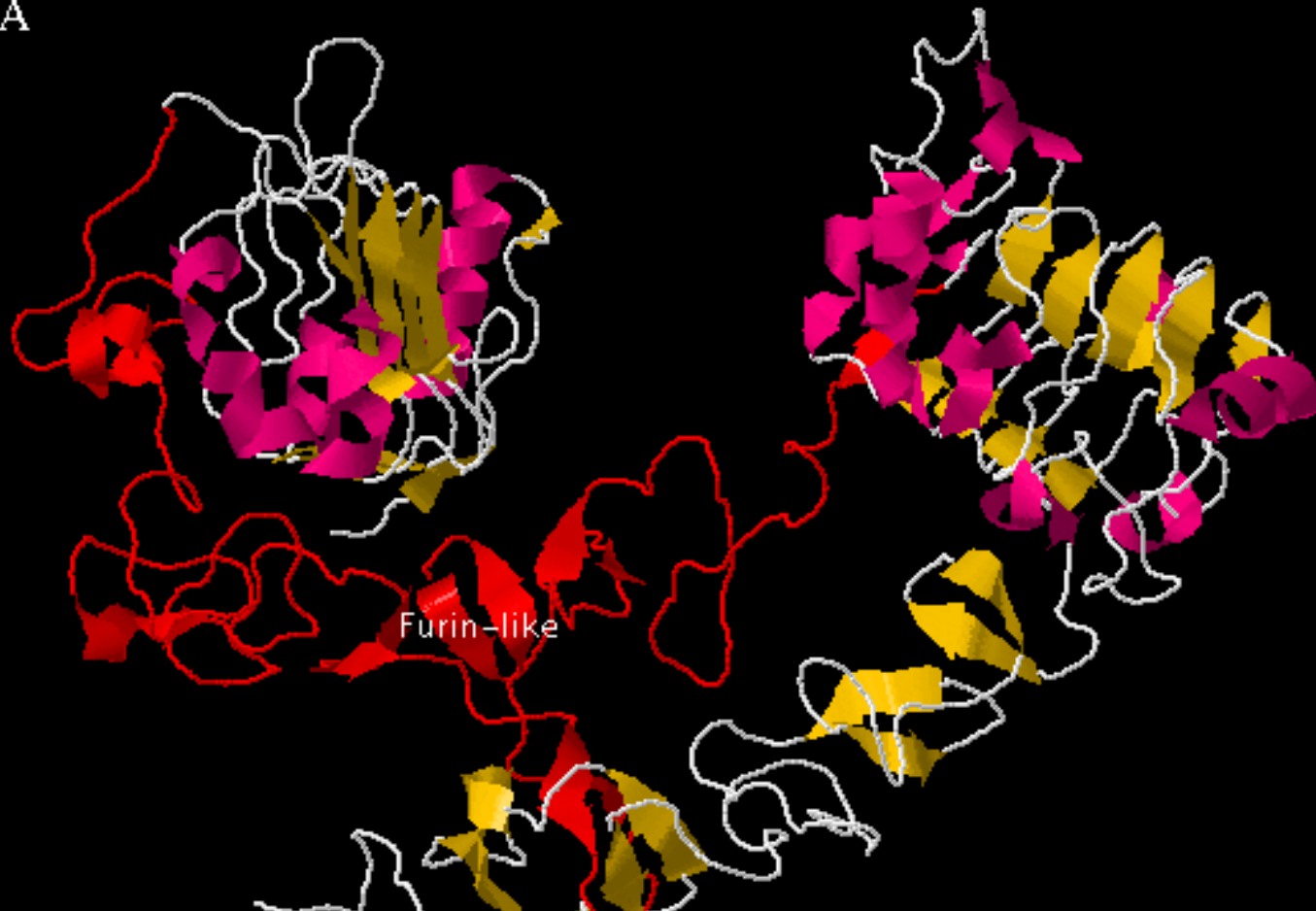
LOCI

HISTORY

Best Annotation ?

UniProtKB/Swiss-Prot:EGFR_HUMAN (Epidermal growth factor receptor precursor (EC 2.7.10.1) (Receptor tyrosine-protein kinase ErbB-1).)

InqlA



Unison Web Tools

unison

UNISON SEARCH BROWSE ANALYZE TOOLS

ALIAN BABELFISH ONTARGET

Query

Enter aliases.

Aliases are identifiers, accessions, or MD5 sequence checksums from any source database.

BCLW
BCL2
BCLX
MCL1

Whitespace and commas will be removed.

submit

Append species identifier.

When the alias does not contain an underscore, append the following species:

HUMAN (Homo sapiens) ▾

Annotation Results

Query	Unison pseq_id	NCBI Gene & RefSeq	GenenGenes	Cytoband	Probes	GO	Domains
BCLW_HUMAN UniProtKB/Swiss-Prot Apoptosis regul...	Unison:30971	GeneID:599 ↗ RefSeq:NP_004041.1 ↗	UNQ34448 ↗, PRO59288 ↗	14q11.2		Function: protein binding Process: anti-apoptosis Process: spermatogenesis	BH4(6-32;52;2.1e-12) BH4_1(9-29) Bcl-2(46-144;192;2.3e-54) BH1(85-104) BH2(137-148) TM(169-191)
BCL2_HUMAN UniProtKB/Swiss-Prot Apoptosis regul...	Unison:30966	GeneID:596 ↗ RefSeq:NP_000624.2 ↗		18q21.33	HGU133P:207004_at ↗ HGU133P:207005_s_at ↗ HGU133P:232210_at ↗ HGU133P:232614_at ↗ HGU133P:237837_at ↗ HGU133P:244035_at ↗	See all 25 functions	BH4(7-33;57;5.6e-14) BH4_1(10-30) BH3(93-107) Bcl-2(97-195;197;5.1e-56) BH1(137-155) BH2(188-199) TM(214-236)
BCLX_HUMAN UniProtKB/Swiss-Prot Apoptosis regul...	Unison:30974	GeneID:598 ↗ RefSeq:NP_612815.1 ↗	UNQ2707 ↗, PRO107805 ↗, PRO116971 ↗, PRO120226 ↗	20q11.21	HGU133P:1569067_at ↗ HGU133P:206665_s_at ↗ HGU133P:215037_s_at ↗	See all 7 functions	BH4(1-27;58;3e-14) BH4_1(4-24) BH3(86-100) Bcl-2(90-188;207;5.2e-59) BH1(130-148) BH2(181-192)

Unison is a platform for diverse tools.

GenenGenes – v3

SEARCH | REQUEST SYSTEMS | ANALYSIS TOOLS | RESOURCES

Quick search: DNA | Enter ic number | qc | new | edit

UNQ2334

Short name:	MCL1
Gene name:	myeloid cell leukemia sequence 1 (BCL2-related)
Gene description:	Myeloid cell leukemia 1, an apoptosis inhibitor, upregulated during cell differentiation, may play roles in multiple myeloma, B-cell non-Hodgkin's lymphomas, chronic lymphocytic leukemia, anaplastic large cell lymphoma and basal cell carcinoma
Gene synonyms:	AW556800,EAT,EAT(TM)/MCL1L/MCL1S/MGC1839/MGC104264, Induced myeloid leukemia cell differentiation protein mcl-1, MCL1,MCL1L,MCL1S,MGC104264, MGC1839,Mcl-1,Mcl1,Mm.1639,TM,myeloid cell leukemia sequence 1, myeloid cell leukemia sequence 1 (BCL2-related), myeloid cell leukemia sequence 1 isoform 1, myeloid cell leukemia sequence 1 isoform 2
Primary DNAs:	Human DNA348250 Mouse DNA188905
UNQ Origin:	GenBank
UNQ type:	Gene
Small Molecule Project:	Bcl-2 Family

[EXPAND TABS]

Protein | **Domain Map** | Genome Map | Family | Gene Summary | Microarray | Publication | Experiment | Inventory

Annotations | Status

eval 0.001 | [Domain Map](#) | [Variant Map](#)

Short Name	Long Name	Num of Domains	Type DNA
Bcl-2	Apoptosis regulator proteins, Bcl-2 family	1	Human
TM	transmembrane domain	1	Human

Matt Brauer
 Guy Cavet
 Josh Kaminker
 Scott Lohr
 Kathryn Woods
 Jean Yuan
 Peng Yue

Unison is a platform for diverse tools.

MVP
MUTATION, VARIANT AND POLYMERASE CHAIN REACTION ANALYSIS

Manage Data | View Data | Project Editor | Help | Home

Gene Report : EGFR

Select Another Gene Go To: --Select--

UNQ ID: 1033
UNQ Shortname: EGFR

Regions Sequenced (66): 500502 501456 504516 504614 504908 505066 505342 506198 506358 506586 506704 506968 507168 507422 508310 508342 509456 510148 510206 510678 510748 510924 511454 511540 512142 512396 515420 515872 516046 516256 516304 516372 516378 516422 516586 516608 516718 650262 e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 e11 e12 e13 e14 e15 e16 e17 e18 e19 e20 e21 e22 e23 e24 e25 e26 e27 e28

Domains: [Recep_L_domain](#) [Furin-like](#) [Recep_L_domain](#) [Pkinase](#) [Pkinase_Tyr](#)

Synonyms: EGFR ERBB ERBB1 UNQ1033 mENA

COSMIC: [108R>K](#) [289A>T](#) [289A>D](#) [289A>V](#) [598G>V](#) [719G>C](#) [719G>A](#) [735G>S](#) [746E>K](#) [746ELREATS>V](#) [...]

Gene Links: [GeneHub](#) [GenenGenes](#) [COSMIC](#) [dbSNP](#)

Gene Graphic

Filter data: Disease Mutation Project SRC Type Somatic

Back to Top Help

1 703 1.405 Kb 2.107 Kb 2.809 Kb 3.511 Kb 4.213 Kb 4.915 Kb

Recep Furin-lik Recep Pkinase_Tyr

e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 e11 e12 e13 e14 e15 e16 e17 e18 e19 e20 e21 e22 e23 e24 e25 e26 e27 e28

Recep_L Furin-like Recep_L d Pkinase_Tyr

Matt Brauer
Guy Cavet
Josh Kaminker
Scott Lohr
Kathryn Woods
Jean Yuan
Peng Yue

Unison is a platform for diverse tools.

MVP

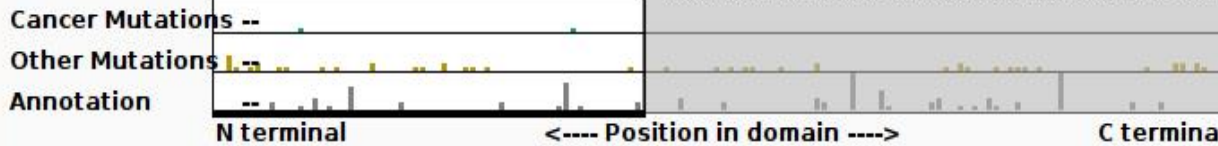
MUTATION, VARIANT AND POLYMERIZATION ANALYSIS
 Manage Data View Data Project Editor Help Home

mCluster

Mutation alignment viewer

Domain Overview: TNF > Show Description

Bar heights reflect the numbers of features at individual alignment positions. Double click each bar to highlight the features in the alignment. Drag the black box to change region shown in views below



Search sequences Search Reset

Help

DB	Species	Gene	Start	End	Residues	AA residue number: 132
Number of Features at Each Position:					41 12 211 2 3111 6 2 2 11 2 11 1 2	1711 12
cosmic	HUMAN	TNF	102	233	LQWLNRANALLANGVELRDNQ-LVVPSEGLYLIYSQVLFK-----GQGC-----ST	
jhu_gbm	HUMAN	TNFSF18.CCDS1367	170	170	SKWQMASSEPPCIN--KVSDWK-LEILQNGLYLIYQGVAPN-----ANYN-----D	
sprot	HUMAN	CD40L	138	261	LQWAEKGYTMSNNLVTLENGKQLTVKRQGLYIYAGVTF-----SNREA-----SS	
sprot	MOUSE	CD40L	137	260	LQWAKKGYTMSNLMLENGKQLTVKREGLYYVYVYVTF-----SNREP-----SS	
sprot	RAT	CD40L	137	260	LQWAKKGYTMSNLMVLENGRQLTVKREGLYYVYVYVTF-----SNREP-----LS	
sprot	HUMAN	CD70	73	191	LYWQGGPAGRSFHLGPELDKGG-LRIHRDGIYMVHIQVTLA-----ICS-----STASR	
sprot	MOUSE	CD70	75	193	LPWGAGPAGRSFTHGPELEEGL-LRIHQDGLYRLHIQVTLA-----NCS-----SPSTL	
sprot	HUMAN	EDA	272	385	NDWSR---ITMNPVKFLHPRSGLEVLVDGTYFIYSQVEV-----	
sprot	MOUSE	EDA	272	385	NDWSR---ITMNPVKFLHPRSGLEVLVDGTYFIYSQVEV-----	
sprot	HUMAN	TN13B	166	284	VPW----LLSFKRGSAALEEKENILVKETGYFFIYGGVLYT-----DK	
sprot	MOUSE	TN13B	190	308	VPW----LLSFKRGNAALEEKENIVVRQTGYFFIYSGVLYT-----DP	
sprot	HUMAN	TNF10	152	280	NSWESSRSGHFSLSNLHLRNGE-LVIHEKGFYYIYSQTYFR-----FQEETI-----KETKN	
sprot	HUMAN	TNF11	185	313	SSWYHDR-GWAKISNMTFSNGK-LIVNQDGFYYLYANICFR-----HHETS-----GDATE	
sprot	MOUSE	TNF11	184	312	SSWYHDR-GWAKISNMTLSNGK-LRVNQDGFYYLYANICFR-----HHETS-----GSPTD	
sprot	RAT	TNF11	186	314	SSWYHDR-GWAKISNMTLSNGK-LRVNQDGFYYLYANICFR-----HHETS-----GSPAD	
sprot	HUMAN	TNF12	131	248	SGWEEARISSSPLRYNRQIGE-FIVTRAGLYYLYCQVHFD-----EGKA-----	
sprot	MOUSE	TNF12	131	248	SGWEEKISSSPLRYDRQIGE-FTVIRAGLYYLYCQVHFD-----EGKA-----	
sprot	HUMAN	TNF13	136	250	VMW----QPALRRGRGLQAQGYVRIQDAGVYLLYSQVLFQ-----DV	
sprot	MOUSE	TNF13	127	241	VMW----QPVLRRGRGLEAQQDVRVWDTGIYLLYSQVLFH-----DV	
sprot	HUMAN	TNF14	112	240	LLWETQL-GLAFLRGLSYHDGA-LVVTKAGYYYIYSKVQLG-----GVGCP-----LGA--	
sprot	MOUSE	TNF14	110	239	LLWETRL-GLAFLRGLTYHDGA-LVTMEPGYYYVYSKVQLS-----GVGCP-----QLAN	

Matt Brauer
 Guy Cavet
 Josh Kaminker
 Scott Lohr
 Kathryn Woods
 Jean Yuan
 Peng Yue

Design Lessons

- **Know what data to integrate, how they'll be used, and the converse.**
- **Integrate on simple, intuitively meaningful abstract concepts.**
 - Precise definitions are critical.
 - Represent proprietary data elsewhere, if needed.
- **Design for Integrity.**
 - Reliability is everything.
- **Aggregate on data types.**
 - Corollary: Partitioning on content makes data silos.

Unison Contents

patents

Geneseq:AAP60074
1991-10-29
SUNTORY
EP205038-A; New tumour...



HUGO

TNFSF9
TNFSF10
TNFSF11

homologs

NP_000585.2	NP_036807.1	RAT
NP_000585.2	NP_038721.1	MOUSE
NP_000585.2	XP_858423.1	CANFA

GO

Function
transcription
initiation
elongation

aliases

TNFA_HUMAN
Q1XHZ6
IPI00001671.1
INCY:1109711.FL1p
CCDS4702.1
gi:25952111



sequences

>Unison:98
MSTESMIRDVE...FGIIAL
>Unison:23782
VRSSSRTPSD...FGIIAL

SNPs

P84L
A94T

Entrez

gene_id
symbol
locus

protein features



1	23		SS
108	143	1.8e-06	EGF
162	184		TM
133	138		ITIM

taxonomy

9606 Homo sapiens
10090 Mus musculus
10028 Rattus rattus



alignments

TNFA 1tnfA
TNFA 1tnfB
...
TNFA 5tswF

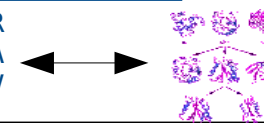
loci

1 233 6+:31651498-31653288



aa-to-resid

MSTESMIR
DVEFGIIA
TESMIRDV
TIAMDAC



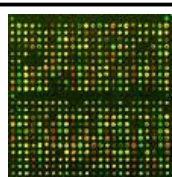
genomes

Hs35
Hs36
RAT



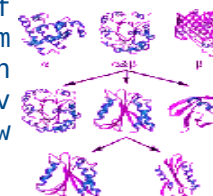
probes

HGU133P
WHG



structures

1tnf
1a8m
2tun
4tsv
5tsw




SCOP

all alpha
all beta
Ig
TNF-like
alpha+beta

Ex1: Mine for sequences w/conserved features.

patents

Geneseq:AAP60074
1991-10-29
SUNTORY
EP205038-A; New tumour...



HUGO

TNFSF9
TNFSF10
TNFSF11

homologs


NP_000585.2	NP_036807.1	RAT
NP_000585.2	NP_038721.1	MOUSE
NP_000585.2	XP_858423.1	CANFA

GO

Function
transcription
initiation
elongation

aliases

TNFA_HUMAN
Q1XHZ6
IPI00001671.1
INCY:1109711.FL1p
CCDS4702.1
gi:25952111



sequences

>Unison:98
MSTESMIRDVE...FGIIAL
>Unison:23782
VRSSSRTPSD...FGIIAL

SNPs

P84L
A94T

Entrez

gene_id
symbol
locus


protein features



1	23	1.8e-06	SS
108	143		EGF
162	184		TM
133	138		ITIM

taxonomy

9606 Homo sapiens
10090 Mus musculus
10028 Rattus rattus



alignments

TNFA 1tnfA
TNFA 1tnfB
...
TNFA 5tswF

loci

1 233 6+:31651498-31653288



aa-to-resid

MSTESMIR
DVEFGIIA
TESMIRDV
TIAMDAC



genomes

Hs35
Hs36
RAT



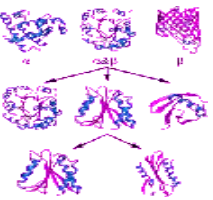
probes

HGU133P
WHG




structures

1tnf
1a8m
2tun
4tsv
5tsw




SCOP

all alpha
all beta
lg
TNF-like
alpha+beta

Ex2: Locate SNPs and Domains on Structure

patents

Geneseq:AAP60074
1991-10-29
SUNTORY
EP205038-A; New tumour...



HUGO

TNFSF9
TNFSF10
TNFSF11

homologs


NP_000585.2	NP_036807.1	RAT
NP_000585.2	NP_038721.1	MOUSE
NP_000585.2	XP_858423.1	CANFA

GO

Function
transcription
initiation
elongation

aliases

TNFA_HUMAN
Q1XHZ6
IPI00001671.1
INCY:1109711.FL1p
CCDS4702.1
gi:25952111



Entrez

gene_id
symbol
locus

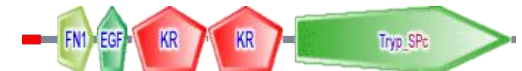
sequences

```
>Unison:98
MSTESMIRDVE...FGIIAL
>Unison:23782
VRSSSRTPSD...FGIIAL
```

SNPs

P84L
A94T


protein features



1	23		SS
108	143	1.8e-06	EGF
162	184		TM
133	138		ITIM

taxonomy

9606 Homo sapiens
10090 Mus musculus
10028 Rattus rattus



loci

1 233 6+:31651498-31653288



alignments

```
TNFA 1tnfA
TNFA 1tnfB
...
TNFA 5tswF
```

aa-to-resid

```
MSTESMIR
DVEFGIIA
TESMIRDV
TIAMDAC
```



genomes

Hs35
Hs36
RAT



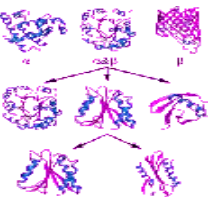
probes

HGU133P
WHG




structures

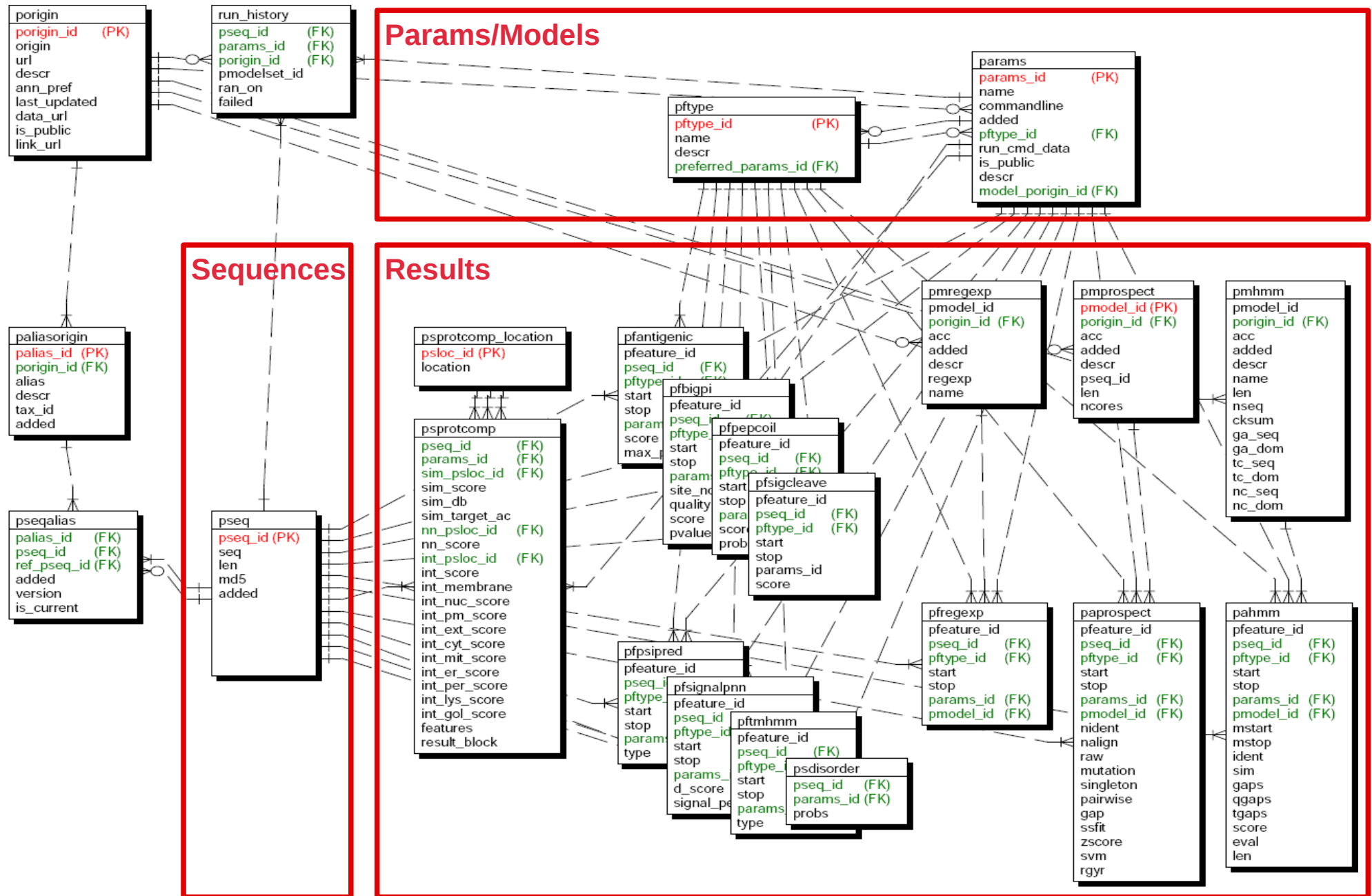
1tnf
1a8m
2tun
4tsv
5tsw



SCOP

all alpha
all beta
lg
TNF-like
alpha+beta

Unison Form Follows Function.



Process Lessons

- **Explicitly track the provenance of data.**
 - All data in Unison are tied to an origin – predictions, annotations, sequences, models.
- **Plan for updates.**
 - Updates are completely automated and idempotent.

idempotent

i·dem·po·tent (/ˈaɪdəmˈpoʊntnt, ˈɪdəm-/)

adj. [from mathematical techspeak] Acting as if used only once, even if used multiple times.

idempotent. Dictionary.com. Jargon File 4.2.0.
<http://dictionary.reference.com/browse/idempotent> (accessed: February 25, 2009).

Unison Build Process



Makefile
downloads all data

Makefile
loads auxiliary data
loads sequences and annotations
(in-house is just another source)
updates sequence sets
updates precomputed predictions
(incremental update!)
updates precomputed analyses and mat'd views

- **Runs in a cron job**
- **Requires ~10% time of 1 person**
- **Consistent, reliable builds**

Other Lessons

- **Design security from the start.**
 - Internal version of Unison use Kerberos.
 - Especially important in a world of distributed services and data.

- **Include web services early in the design.**

Kiran Mukhyala

Fernando Bazan, Matt Brauer, David Cavanaugh, Jason Hackney, Pete Haverty, Ken Jung, Josh Kaminker, Nandini Krishnamurthy, Li Li, Yun Li, Scott Lohr, Shih-ming Loh, Jinfeng Liu, Peng Yue, Jianjun Zhang, Yan Zhang

Simran Hansrai, Marc Lambert, Dave Windgassen

<http://unison-db.org/>

Open access web site, downloads, documentation, references, credits.

unison-db.org:5432

PostgreSQL & odbc/jdbc/sdbc access



“Are you sure about this Stan? It seems odd that a pointy head and a long beak is what makes them fly.”

J. Workman, Science 245:1399 (1989)

Unison facilitates complex mining.

Functional characterization of the *Bcl-2* gene family in the zebrafish

E Kratz¹, PM
R Hart² and

¹ Department of
CA, USA
² Department of
³ Department of
⁴ The Walter and
Australia
* Corresponding
Genentech Inc
Tel: 650-225-

Received 12.6.08
Edited by G Meir

Abstract

Members of
apoptosis pa
in vertebrate
(*Danio rerio*)
structural an
family memb
proteins and
-irradiation
prevented b
Bcl-2 family
was homolo
synteny, ye
human *Bak*.
zMc1-1b rev
zebrafish de
activation th
substantial
mammalian
as a releva
pathway.
Cell Death ar
doi:10.1038/s

Cell Death and Differentiation (2006) 13, 1631–1640
© 2006 Nature Publishing Group All rights reserved 1350-9047/06 \$30.00
www.nature.com/cdd



Genome Biology
IMPACT FACTOR 7.17

Log on / register

BioMed Central home | Journals A-Z | Feedback | Support

home | comment | reviews | reports | deposited research | refereed research | interactions | supplements | search | information | my journal

→ refereed research

refereed research
software | methods | about refereed | upload! | feedback
research
sort by date
sort by subject
list all

Research

Highly accessed Open Access

Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective

Jinfeng Liu ✉, Yan Zhang ✉, Xingye Lei ✉ and Zemin Zhang ✉

Genome Biology 2008, 9:R69 doi:10.1186/gb-2008-9-4-r69

Published: 8 April 2008

Abstract (provisional)

Background

The rates of molecular evolution for protein-coding genes depend on the stringency of functional or structural constraints. The Ka/Ks ratio has been commonly used as an indicator of selective constraints and is typically calculated from interspecies alignments. Recent accumulation of single nucleotide polymorphism data has enabled the derivation of Ka/Ks for polymorphism (SNP A/S ratio).

Results

Using data from dbSNP, we conducted the first large-scale survey of SNP A/S ratios for different structural and functional properties. We confirmed that the SNP A/S ratio is largely correlated with Ka/Ks for divergence. We observed stronger selective constraints for proteins that have high mRNA expression level or broad expression patterns, have no paralogs, arise earlier in evolution, have natively disordered region, are located in cytoplasm and nucleus, or are related to human diseases. On the residue level, we found higher degree of variation for residues that are exposed to solvent, are in loop conformation, natively disordered regions or low complexity regions, or are in the signal peptides of secreted proteins. Our analysis also revealed that histones and protein kinases are among the protein families that are under the strongest selective constraints, whereas olfactory and taste receptors are among the most variable groups.

Conclusions

Our study suggests that the SNP A/S ratio is a robust measure for selective constraints. The correlations between SNP A/S ratios and other variables provide valuable insights into the natural selection of various structural or functional properties, particularly for human-specific genes and constraints within the human lineage.

Genome Biology

Volume 9
Issue 4

Viewing options:

- Abstract
- PDF (304KB)

Associated material:

- PubMed record

Related literature:

- Articles citing this article on PubMed Central
- Other articles by authors on Google Scholar on PubMed
- Related articles/pages on Google on Google Scholar on PubMed

Tools:

- Email to a friend
- Order reprints
- Post a comment
- Sign up for article alerts

Post to:

- Citeulike
- Connotea
- Del.icio.us
- Digg
- Facebook

Jason Hackney
Nandini Krishnamurthy
Li Li
Yun Li
Jinfeng Liu
Shiu-ming Loh
Kiran Mukhyala