Unison: Enabling easy, rapid, and comprehensive proteomic mining

http://unison-db.org/
Online access, download, documentation, references.

Reece Hart
Genentech, Inc.

UCSF / SF PostgreSQL Users' Group
March 11, 2009
San Francisco, CA

Slides 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, networks ...

**Chemistry**
compounds, HCS, HTS, properties ...

**Annotation**
GO, taxonomy, SCOP, disease, OMIM ...

**Clinical**
assays, protocols, patient records, samples ...

**LIMS**
animal records, protocols request systems, personnel, samples ...

**Communications**
literature, patents, and presentations ...


➢ **Source Aggregation**
- Aggregates data of the same type from multiple sources.
- Ensures completeness of data.

➢ **Semantic Integration**
- Integrates fundamentally distinct data types.
- Abstracts types to essential features.
- Improves contextual understanding of data.
A Survey of Integration Methods

Presentation

Link Integration
Hypertext links between sites

Mashups
AJAX, iframe

Middle Tier

Server Mashups

Database Integration
(Federation / Warehouse)

Federation

Warehouse

Source Databases
or Files

For review, see:
Goble C, Stevens R
The Problems in a Nutshell

- **Data integration is complex.**
  - Establishing semantic equivalences and relationships are difficult.
  - Source database contents are updated often.

- **Existing tools don't cut it.**
  - Licensing restrictions prevent sharing.
  - Narrow in type of data and/or content, and not easily updated.
  - Not specifically designed for mining.

- **Scientists develop ad hoc and integration solutions.**
  - Results are difficult to repeat.
  - It wastes a lot of time.
  - Questions don't get asked.
Unison in a Nutshell

Protein Sequences and Annotations
Genomes, Gene Mapping & Structure, Probes
Domain, Structure & Homology Predictions
Structures & Ligands
Auxiliary Annotations
HomoloGene, Gene Ontology, taxonomy, PDB, HUGO, SCOP, etc.
GO, RIF, SCOP, etc.

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

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

**Entrez**
gene_id
symbol
locus

**taxonomy**
9606 Homo sapiens
10090 Mus musculus
10028 Rattus rattus...

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

**SNPs**
P84L
A94T

**alleles**
TNFA HUMAN
Q1XHZ6
IPI00001671.1
INCY:1109711.Fl1p
CCDS4702.1
gi:25952111

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

**loci**
1 233 6+:31651498-31653288

**alignments**
TNFA 1tnfA
TNFA 1tnfB
...
TNFA 5tswF

**structures**
1tnf
1a8m
2tun
4tsw
5tsw

**aa-to-resid**
MSTESMIR
DVEFGIIA
TESMIRDV
IIIAMDAC

**probes**
HGU133P
WHG

**scop**
all alpha
all beta
Ig
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

**sequences**
>Unison: 98
MSTESMIRDVE...FGIIAL
>Unison: 23782
VRSSRTPSD...FGIIAL

**alignments**
TNFA 1tnfA
TNFA 1tnfB
...TNFA 5tswF

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

**structures**
1tnf
1a8m
2tun
4tsv
5tsw

**aa-to-resid**
MSTESMIR
DVEFGIIA
TESMIRDV
IIAMDAC

**alias**
TNFA HUMAN
Q1XZH6
IPI00001671.1
INCY:1109711.FL1p
CCDS4702.1
gi:25952111

**taxonomy**
9606 Homo sapiens
10090 Mus musculus
10028 Rattus rattus

**Entrez**
gene_id
symbol
locus

**GO**
Function
transcription
initiation
elongation

**Genomes**
Hs35
Hs36
RAT

**probes**
HGU133P
WHG

**SCOP**
all alpha
all beta
Ig
TNF-like
alpha+beta
Analysis and data mining have distinct needs.

**Feature-based Mining**
- i.e., show sequences that contain specified features
- Typically entails days to months of computing results.

**Sequence Analysis**
- i.e., show predictions for a given sequence
- Typically involves minutes to hours of computing per sequence.

**Prediction results**
- Method-specific data such as score, e-value, p-value, kinase probability, etc.

**Parameters**
- Execution arguments/options for every prediction type and result

**Feature types/models**
- HMM, TM, signal, etc.

**Source integration**
- Sequences non-redundant superset of all sequences

**Semantic integration**
- Feature types/models
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.

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 pfregexp_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';

<table>
<thead>
<tr>
<th>pseq_id</th>
<th>Ig_start</th>
<th>Ig_stop</th>
<th>score</th>
<th>eval</th>
<th>TM_start</th>
<th>Tm_stop</th>
<th>ITIM_start</th>
<th>ITIM_stop</th>
<th>best_annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>262</td>
<td>316</td>
<td>30</td>
<td>7.40E-06</td>
<td>440</td>
<td>462</td>
<td>518</td>
<td>523</td>
<td>UniProtKB/Swiss-Prot:SIGL5_HUMAN (Rec</td>
</tr>
<tr>
<td>254</td>
<td>158</td>
<td>213</td>
<td>36</td>
<td>1.90E-07</td>
<td>284</td>
<td>306</td>
<td>386</td>
<td>391</td>
<td>UniProtKB/Swiss-Prot:VSIG4_HUMAN (Rec</td>
</tr>
<tr>
<td>544</td>
<td>157</td>
<td>215</td>
<td>24</td>
<td>6.60E-04</td>
<td>348</td>
<td>370</td>
<td>431</td>
<td>436</td>
<td>UniProtKB/Swiss-Prot:SIGL9_HUMAN (Rec</td>
</tr>
<tr>
<td>797</td>
<td>254</td>
<td>312</td>
<td>40</td>
<td>7.60E-09</td>
<td>1099</td>
<td>1121</td>
<td>1361</td>
<td>1366</td>
<td>UniProtKB/Swiss-Prot:DCC_HUMAN (Rec</td>
</tr>
<tr>
<td>1113</td>
<td>42</td>
<td>102</td>
<td>30</td>
<td>1.20E-05</td>
<td>243</td>
<td>265</td>
<td>300</td>
<td>305</td>
<td>UniProtKB/Swiss-Prot:KI2L2_HUMAN (Re</td>
</tr>
<tr>
<td>1114</td>
<td>42</td>
<td>102</td>
<td>30</td>
<td>6.50E-06</td>
<td>243</td>
<td>265</td>
<td>330</td>
<td>335</td>
<td>UniProtKB/Swiss-Prot:KI2L1_HUMAN (Re</td>
</tr>
<tr>
<td>1115</td>
<td>42</td>
<td>102</td>
<td>31</td>
<td>4.20E-06</td>
<td>243</td>
<td>265</td>
<td>301</td>
<td>306</td>
<td>UniProtKB/Swiss-Prot:KI2L3_HUMAN (Re</td>
</tr>
<tr>
<td>1116</td>
<td>42</td>
<td>97</td>
<td>30</td>
<td>1.10E-05</td>
<td>339</td>
<td>361</td>
<td>396</td>
<td>401</td>
<td>UniProtKB/TrEMBL:Q95368_HUMAN (Sub</td>
</tr>
<tr>
<td>1134</td>
<td>340</td>
<td>388</td>
<td>26</td>
<td>1.40E-04</td>
<td>603</td>
<td>625</td>
<td>688</td>
<td>693</td>
<td>UniProtKB/Swiss-Prot:PECA1_HUMAN (Re</td>
</tr>
</tbody>
</table>
Unison has many applications.

**Unison Web Tools**
- Protein Sequences and Annotations
- Genomes, Gene Mapping & Structure, Probes
- Domain, Structure & Homology Predictions

**Other In-House Tools**
- Structures & Ligands
- Auxiliary Data
- Precomputed Predictions

**Ad Hoc Mining**
- Mining and analysis projects

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

E. Krätz, P. M. Eimon, K. Mukhyala, H. Stern, J. Zha, A. Strasser, R. Hart, and A. Ashkenazi.

Department of Molecular Oncology, Genentech Inc., South San Francisco, CA, USA.
Department of Bioinformatics, Genentech Inc., South San Francisco, CA, USA.
Department of Pathology, Genentech Inc., South San Francisco, CA, USA.

Corresponding author: A. Ashkenazi, Department of Molecular Oncology, Genentech Inc., 1 DNA Way, MS 42, South San Francisco, CA 94080, USA. Email: aashkenazi@gene.com

Received 12/6/06; accepted 26/6/06; published online 4/8/06

Edited by G. Malini

Abstract

Members of the Bcl-2 protein family control the intrinsic apoptosis pathway. To evaluate the importance of this family in vertebrate development, we investigated it in the zebrafish (Danio rerio). We found that the zebrafish genome encodes structural and functional homologs of most mammalian Bcl-2 family members, including multi-Bcl-2 homology (BH) domain proteins and BH3-only proteins. Apoptosis induction by γ-irradiation required zBax1 and zPuma, and could be prevented by overexpression of homologs of prosurvival Bcl-2 family members. Surprisingly, zebrafish Bax2 (zBax2) was homologous to mammalian Bax by sequence and stynteny, yet demonstrated functional conservation with human Bak. Morpholino knockdown of both zBcl-1a and zMcl-1b revealed their critical role in early embryonic zebrafish development, and in the modulation of apoptosis activation through the extrinsic pathway. These data indicate substantial functional similarity between zebrafish and mammalian Bcl-2 family members, and establish the zebrafish as a relevant model for studying the intrinsic apoptosis pathway.

http://dx.doi.org/10.1080/09678820600966166; published online 4 August 2006

three classes of proteins: prosurvival (including Bcl-2, Bcl-xL, Mcl-1, A1, Bcl/BI, Bcl-B, and Bcl-w), multidomain proapoptotic (including Bax, Bak, and Bok), and proapoptotic BH3 only (Bid, Bad, Bmf, Bik, Puma, Noxa, Hrk, and BNIP3). In unstimulated cells, interactions between prosurvival and proapoptotic multidomain family members prevent the latter from oligomerizing and initiating the apoptotic program. Upon apoptotic stimulation, BH3-only proteins relieve the inhibition of multidomain proapoptotic proteins, freeing them to initiate apoptosis. The subsequent release of proapoptotic factors from the mitochondria results in the stimulation of caspase-9, and cleavage and activation of effector caspases such as caspase-3, -6, and -7.

Previous experiments demonstrate that several members of the Bcl-2 family are required for normal development; however, the function of this protein family during development is largely unknown. Simultaneous knockdown of Bax and Bak results in drastically decreased perinatal survival, with fewer than 10% of the double-knockout mice surviving to adulthood. These mice are born in normal Mendelian ratios, but are neglected by the mother and likely die from failure to nurse, perhaps as a result of massive inappropriate neuronal accumulation. Bcl-x knockout mice display a critical role for Bcl-x in the survival of developing neuronal and hematopoietic cell types, whereas the Bcl-2 knockout has only moderate effects in adult mice. Knockout of Mcl-1 results in preimplantation lethality. While these experiments clearly indicate important roles for Bcl-2 family members in early development, in each of these knockouts neither the initiating apoptotic signal nor the activated BH3-only proteins are known.

We have therefore characterized the complement of Bcl-2 family proteins in zebrafish. This genetically tractable model organism offers the unique opportunity to elucidate the role of the Bcl-2 genes during embryogenesis ex utero. While the existence of several Bcl-2 family members in zebrafish has been reported, little is known about the functional homology between zebrafish and mammalian Bcl-2 family members. Here, we report that the zebrafish genome contains full-length, functional homologs of most Bcl-2 family members. Overexpression of homologs of most mammalian

Unison facilitates complex mining.

Jason Hackney
Nandini Krishnamurthy
Li Li
Yun Li
Jinfeng Liu
Shiu-ming Loh
Kiran Mukhyala
Data integration led to Bcl-2 discoveries.

Custom model building

<table>
<thead>
<tr>
<th>Z'fish Protein</th>
<th>Source Database and Accession</th>
<th>Human Protein</th>
<th>E-value</th>
<th>Score</th>
<th>% Ide</th>
<th>% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bax</td>
<td>RefSeq:NP_571637</td>
<td>BAX</td>
<td>2.00E-47</td>
<td>189</td>
<td>51</td>
<td>98</td>
</tr>
<tr>
<td><strong>Bax2</strong></td>
<td>E35:ENSDARP00000040899</td>
<td>BAX</td>
<td>1.00E-14</td>
<td>81</td>
<td>33</td>
<td>51</td>
</tr>
<tr>
<td><strong>Bik</strong></td>
<td>UP:Q5RGV6_BRARE</td>
<td>BIK</td>
<td>1.41E+04</td>
<td>20</td>
<td>47</td>
<td>51</td>
</tr>
<tr>
<td>Bmf</td>
<td>RefSeq:NP_001038689</td>
<td>Bmf</td>
<td>1.00E-05</td>
<td>50</td>
<td>32</td>
<td>91</td>
</tr>
<tr>
<td><strong>Bmf2</strong></td>
<td>E35:FGENESH00000082230</td>
<td>PUMA</td>
<td>1.10E-02</td>
<td>42</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td><strong>BBC3</strong></td>
<td>E35:FGENESH00000078270</td>
<td>BCL2A1</td>
<td>2.10E+01</td>
<td>30</td>
<td>25</td>
<td>49</td>
</tr>
</tbody>
</table>

4 novel Bcl-2 proteins in zebrafish

### Unison Web Tools

#### Query
Enter aliases. Aliases are identifiers, accessions, or MD5 sequence checksums from any source database.

<table>
<thead>
<tr>
<th>Query</th>
<th>Unison</th>
<th>NCBI Gene &amp; RefSeq</th>
<th>GenenGenes</th>
<th>Cytoband</th>
<th>Probes</th>
<th>GO</th>
<th>Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCL2_HUMAN</td>
<td>Unison:30966</td>
<td>GenID:596 RefSeq:NP_00624.2</td>
<td></td>
<td>18q21.33</td>
<td>HGU1.33P:207004 at, HGU1.33P:207005 s at, HGU1.33P:232210 at, HGU1.33P:237837 at, HGU1.33P:244035 at</td>
<td>See all 25 functions</td>
<td>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)</td>
</tr>
<tr>
<td>BCLX_HUMAN</td>
<td>Unison:30974</td>
<td>GenID:598 RefSeq:NP_612815.1</td>
<td>UN02707, PRC107805, PRC116971, PRC120226</td>
<td>20q11.21</td>
<td>HGU1.33P:1569067 at, HGU1.33P:206665 s at, HGU1.33P:215037 s at</td>
<td>See all 7 functions</td>
<td>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)</td>
</tr>
</tbody>
</table>

### Annotation Results

Whitespace and commas will be removed.

Submit
Structure Viewer with User Features!

http://unison-db.org/pseq_structure.pl?q=TNFA_HUMAN;userfeatures=Estrand@164-174,mysnp@170

Structures/Templates (top 5 hits of 8, including modeled structures)

1a8mB
2tnfB
2tnfA
1tnfC
2e7aB

HMM (hmmer); 1 w/eval<=1

TNF; S=210; E=5.8e-60

SNPs

P->L
A->T

User Features

Estrand
mysnp
Unison is a platform for diverse tools.
## Unison Build Process

<table>
<thead>
<tr>
<th>Phase 0</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Phase 4</th>
<th>Phase 5</th>
<th>Phase 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download</td>
<td>Load Aux Data</td>
<td>Load Sequences</td>
<td>Update Sets</td>
<td>Update Predictions</td>
<td>Update Mat Views</td>
<td>Housekeeping</td>
</tr>
<tr>
<td>2 d</td>
<td>4 h</td>
<td>2 d</td>
<td>1 h</td>
<td>50 CPU-d</td>
<td>6 h</td>
<td>0</td>
</tr>
</tbody>
</table>

- 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
- builds public database

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

➢ Aggregate on data types.
   • Corollary: Partitioning on content makes data silos.

➢ Design for Integrity.
   • Reliability is everything.
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 (ˈaɪdəmˈpou̇tnt, ˈɪdəm-) adj. [from mathematical techspeak] Acting as if used only once, even if used multiple times.

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.
  • (Ooops, I blew it on this.)
A Few Reasons for PostgreSQL.

- Excellent support for server-side functions
  - in PL/PGSQL, Perl, C, Java, Python, R, sh, ...
- Table inheritance
  - Facilitates type abstraction
- GSSAPI/Kerberos support
  - No password admin
  - User identity all the way to the database
- psql rocks
- Pedantic and responsive development community
- Ease community adoption (?)
“Are you sure about this Stan? It seems odd that a pointy head and a long beak is what makes them fly.”


Kiran Mukhyala

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

Simran Hansrai, Marc Lambert, Dave Windgassen

A huge open science and open source community.

http://unison-db.org/
Open access web site, downloads, documentation, references, credits.

unison-db.org:5432
PostgreSQL & odbc/jdbc/sjdbc access
Unison form follows function.