

Reece Hart <rk@gene.com>, Kiran Mukhyala ♦ Departments of Bioinformatics and Protein Engineering ♦ Genentech, Inc. ♦ San Francisco, CA 94080

**Abstract**  
We have recently developed Unison, an Open Source database of extensive precomputed proteomic predictions, and have applied this tool to several function prediction and mining endeavors. Unison has been an indispensable tool in identifying and eliminating mining candidates for tumor necrosis factor ligands, helical cytokines, and death fold proteins, and has provided the foundation for numerous analysis studies. In this poster, we give an overview of Unison and provide an example of its use in mining for Immunoreceptor Tyrosine Inhibitory Motif (ITIM)-containing proteins and Tumor Necrosis Factor ligands, and in analyzing NOD2 polymorphisms.

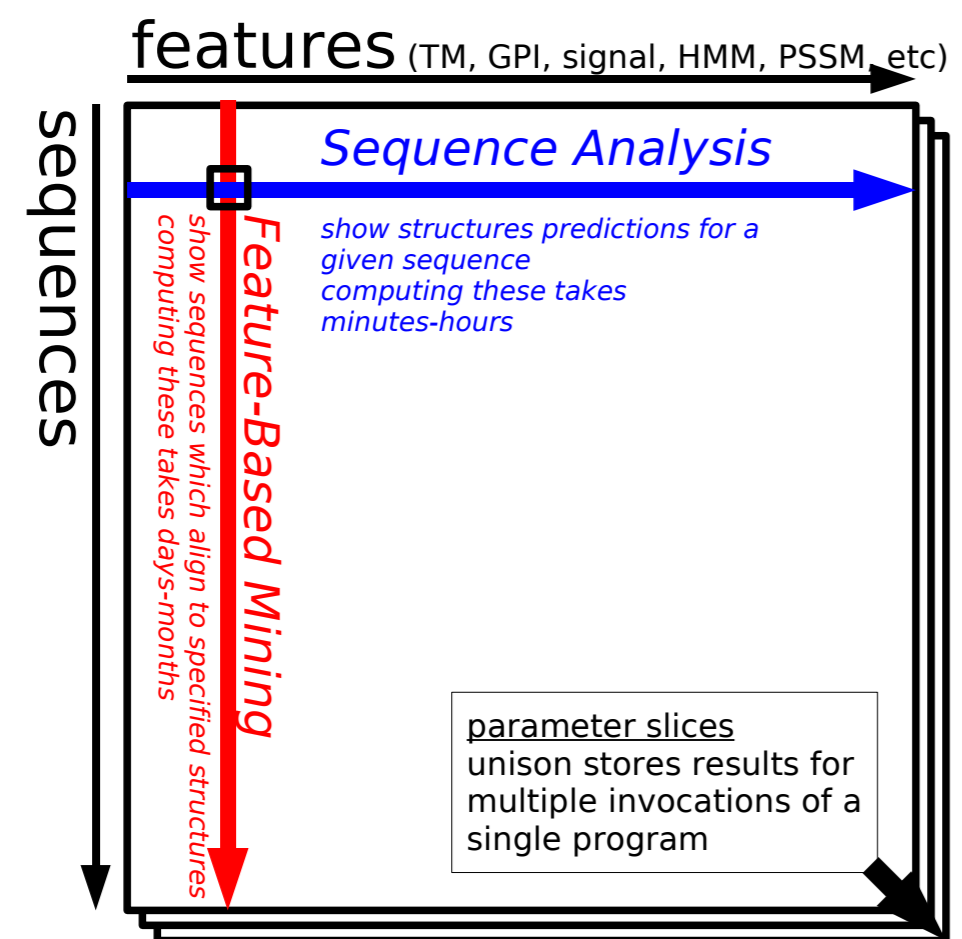
**Motivation and Design**  
Unison is a hypothesis: *that providing a vast universe of non-redundant sequences and a large body of precomputed features on those sequences will enable Genentech to identify or dismiss therapeutic targets more efficiently than with conventional computational pipeline approaches.*

- Unison was motivated by four problems faced during extensive mining efforts:
- Sequence redundancy (exact identity) within and among source databases is an enormous waste of computational and user analysis time.
  - Searching flat files for predictions results was slow, and integrating the results of multiple searches was prohibitively slow.
  - Integrating analyses with external information (e.g., structure classifications, orthologs, or SNPs) was tedious.
  - Keeping predictions and analyses up-to-date with source databases (sequence or model databases) was onerous.

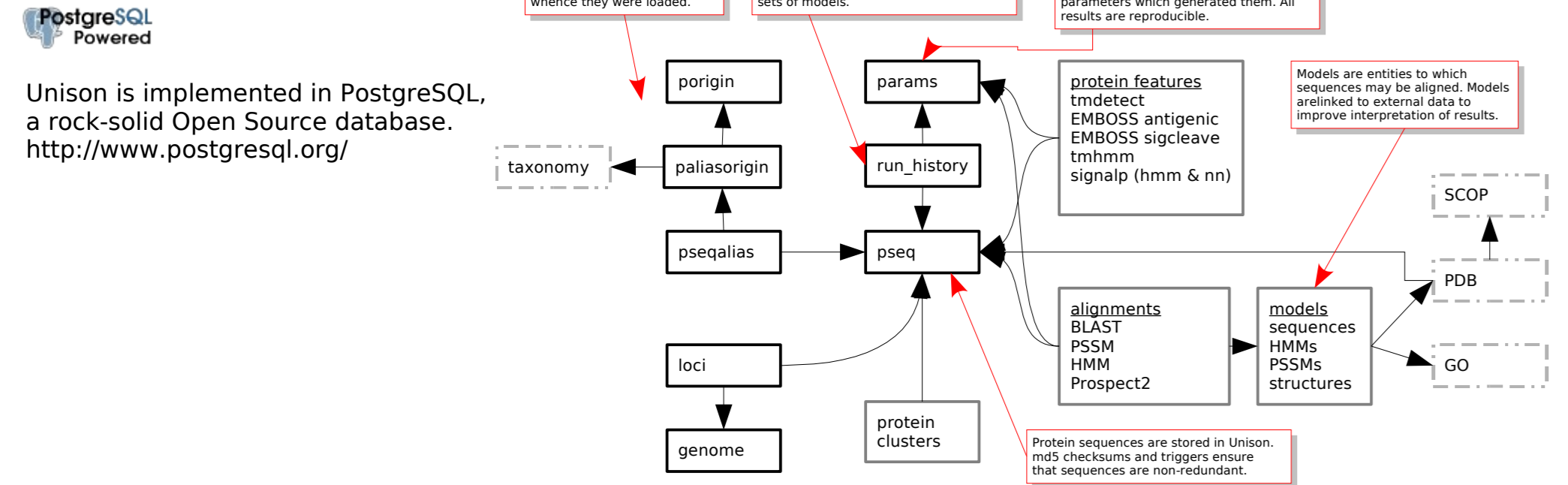
**Sequence Analysis v. Feature-Based Mining**

Sequence analysis: Given a sequence, show features. This typically takes minutes to hours for all features.

Feature-based mining: Given features, identify matching sequences. Computing selected features for all sequences could easily take days to months.



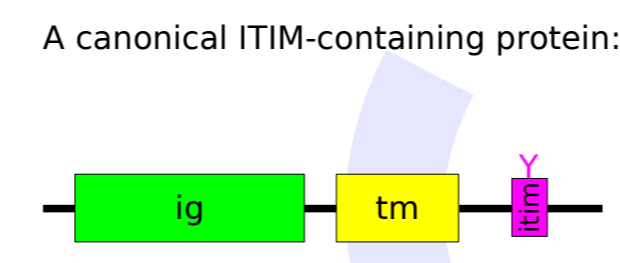
**Simplified Unison Schema**



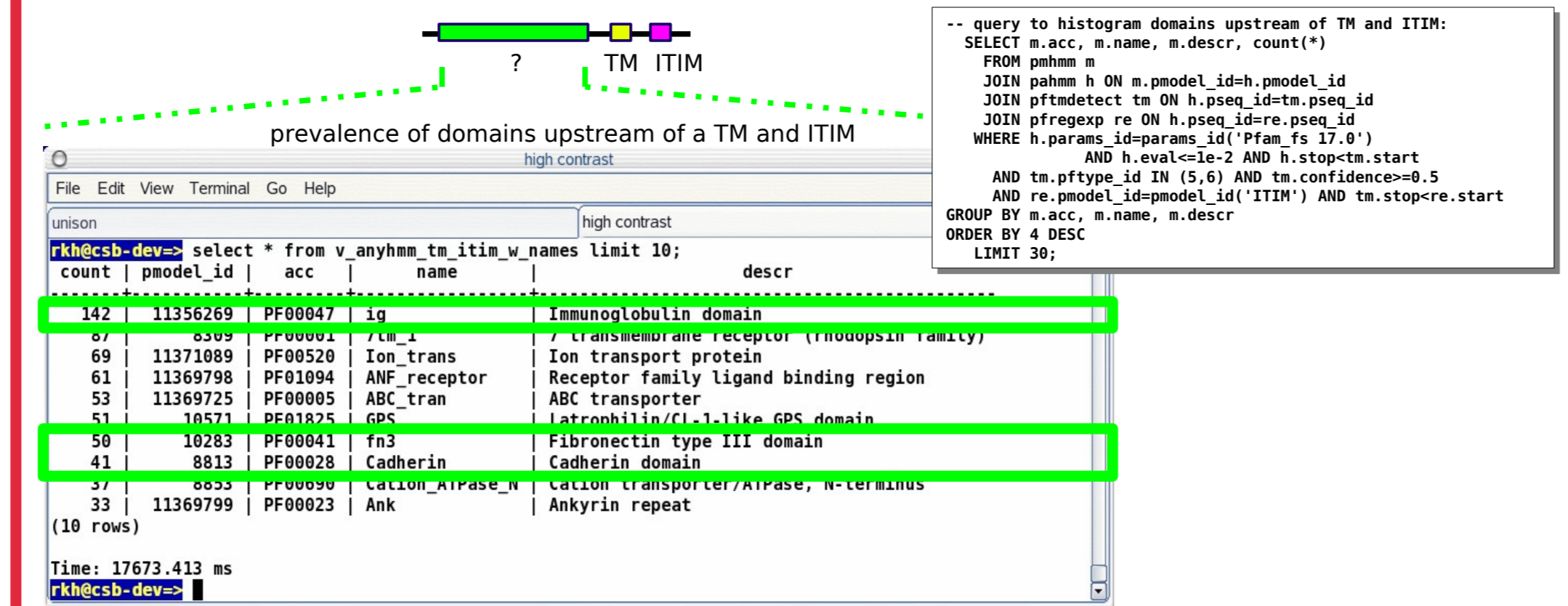
**Mining for Immunoreceptor Tyrosine Inhibitory Motif (ITIM) Proteins**  
ITIMs are short motifs (6-8 AA) in the intracellular domain of immune receptors. Phosphorylation of the tyrosine within an ITIM leads to the binding of SHP1 or SHP2 phosphatases and the attenuation of a corresponding activating receptor. ITIM receptors were initially identified on the surface of NK cells and macrophages where they are believed to prevent self reactivity.

1. Can we identify known ITIMs with a SQL query?  
The following query, modeled after the canonical ITIM shown at right, recovers 36 of 37 known ITIMs (per Staub et al. Cell. Sig., 2003; 1 "known" has since been recalled from RefSeq).

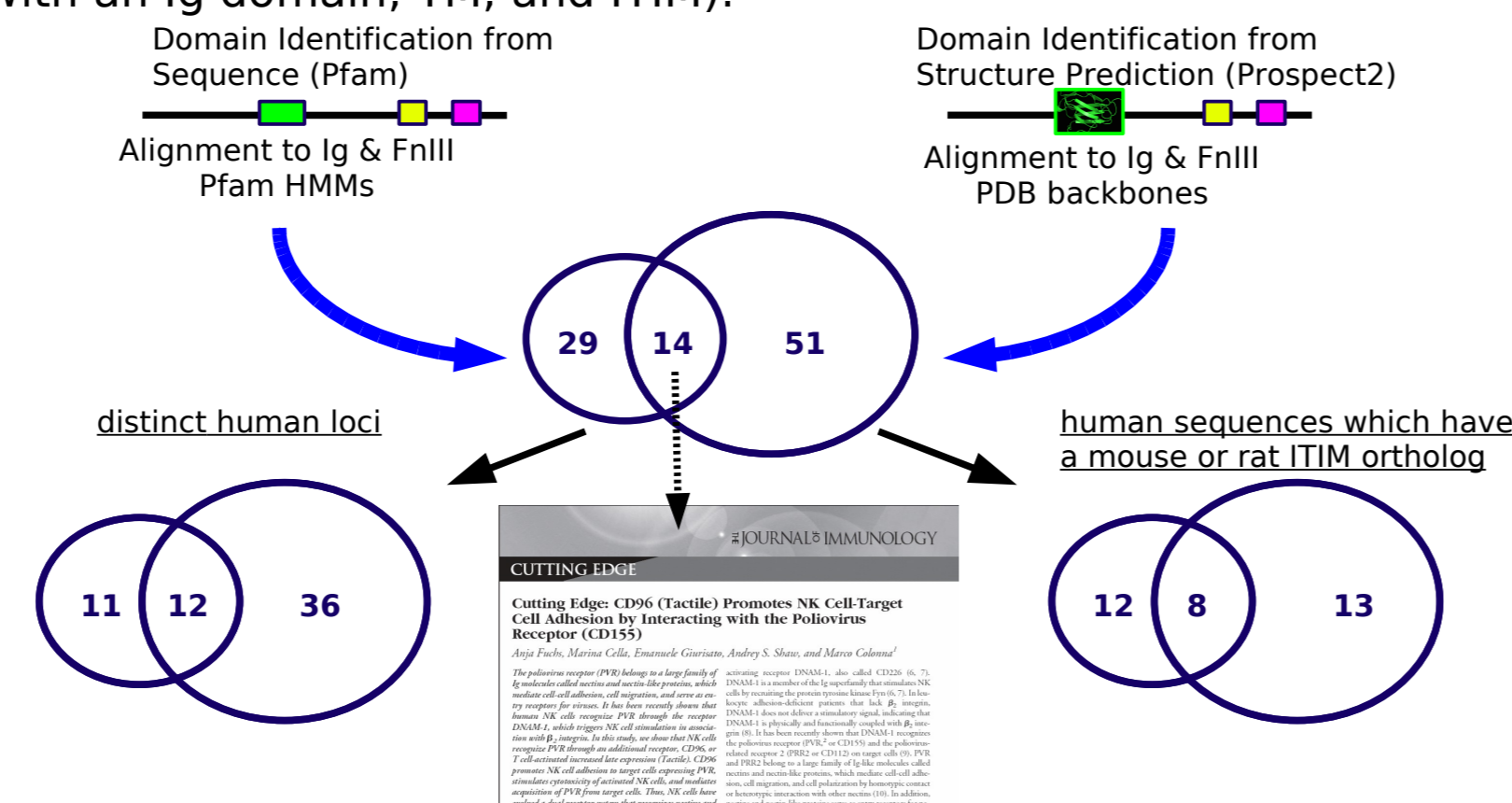
```
SELECT
  hmm_pseq_id,
  hmm_start as 'ig_start',hmm_stop as 'ig_stop',
  tm_start as 'tm_start',tm_stop as 'tm_stop',
  re_start as 'ITIM_start',re_stop as 'ITIM_stop'
FROM
  pfhmm hmm
JOIN pftdetect tm on hmm_pseq_id = tm_pseq_id
JOIN pfrexp re on hmm_pseq_id = re_pseq_id
WHERE
  hmm_pmodel_id=pmodel_id('ig') AND hmm_eval <= 0.02 AND hmm_params_id=params_id('Pfam fs 14.0')
  AND tm_pftype_id in (5,6) AND tm_start > ig_stop AND tm_confidence >= 0.5
  AND re_pmodel_id=pmodel_id('ITIM') AND re_start > tm_stop
```



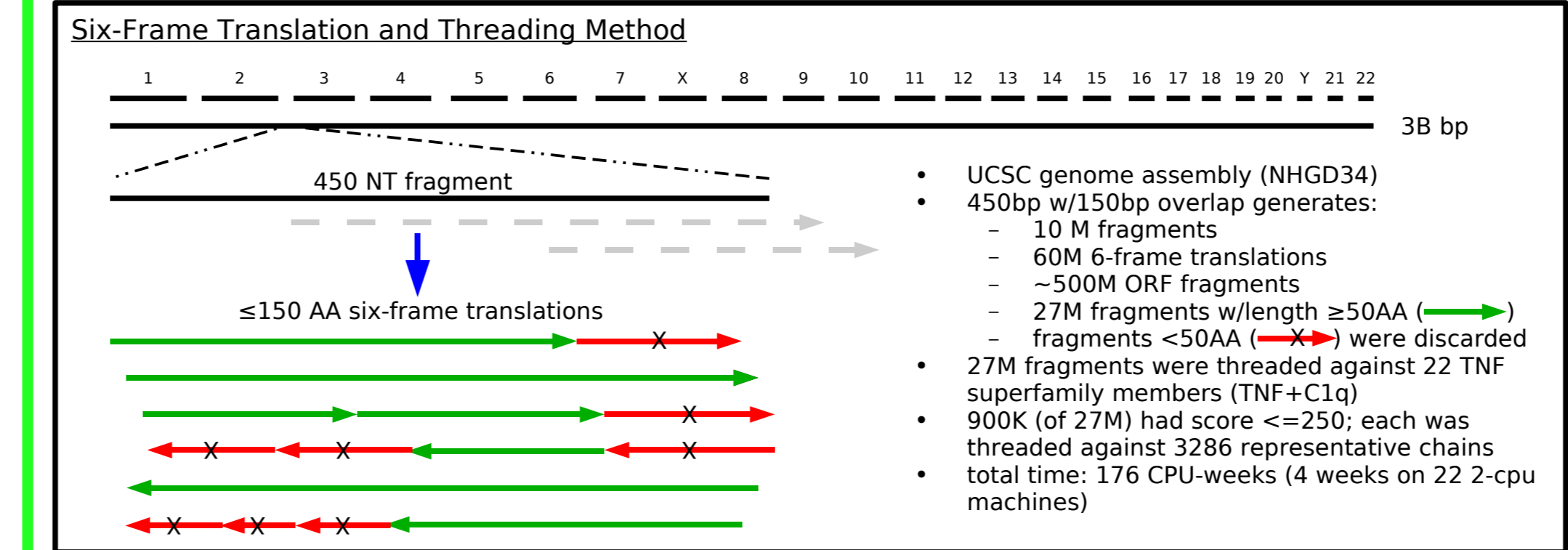
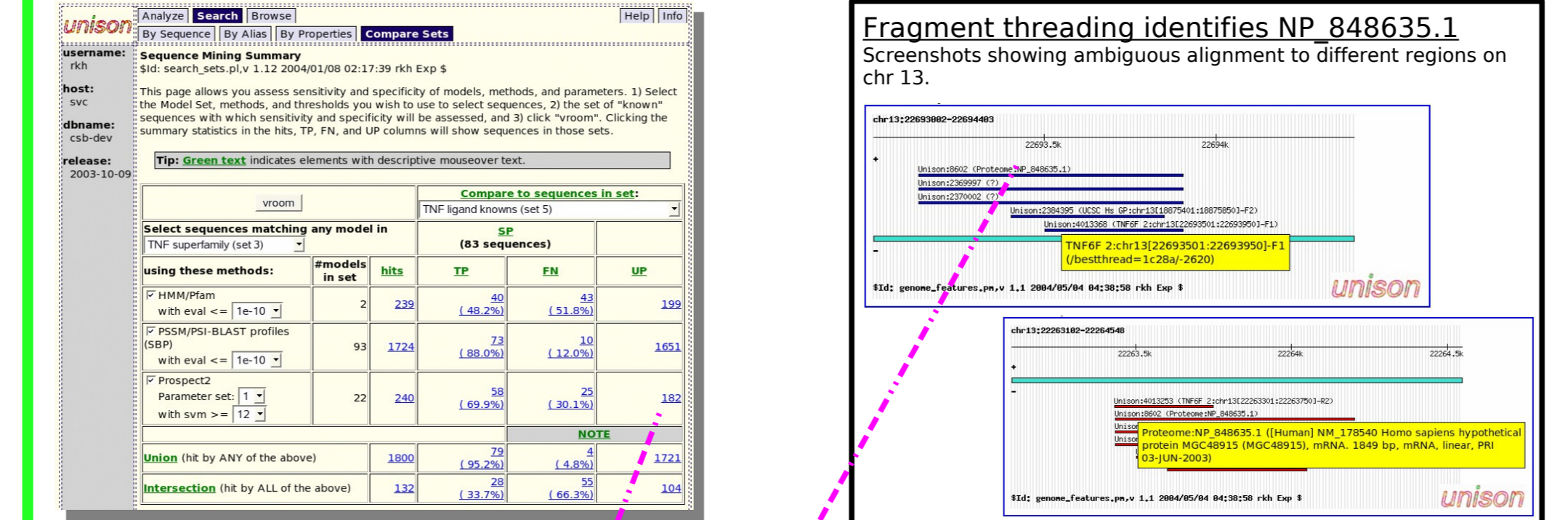
2. What other domains occur in the ECD of an ITIM containing protein?  
Unison may be used to generate *in silico* hypotheses, such as proposing that ITIMs might utilize other 7-sheet beta sandwich folds in their ECD (or other domains entirely).



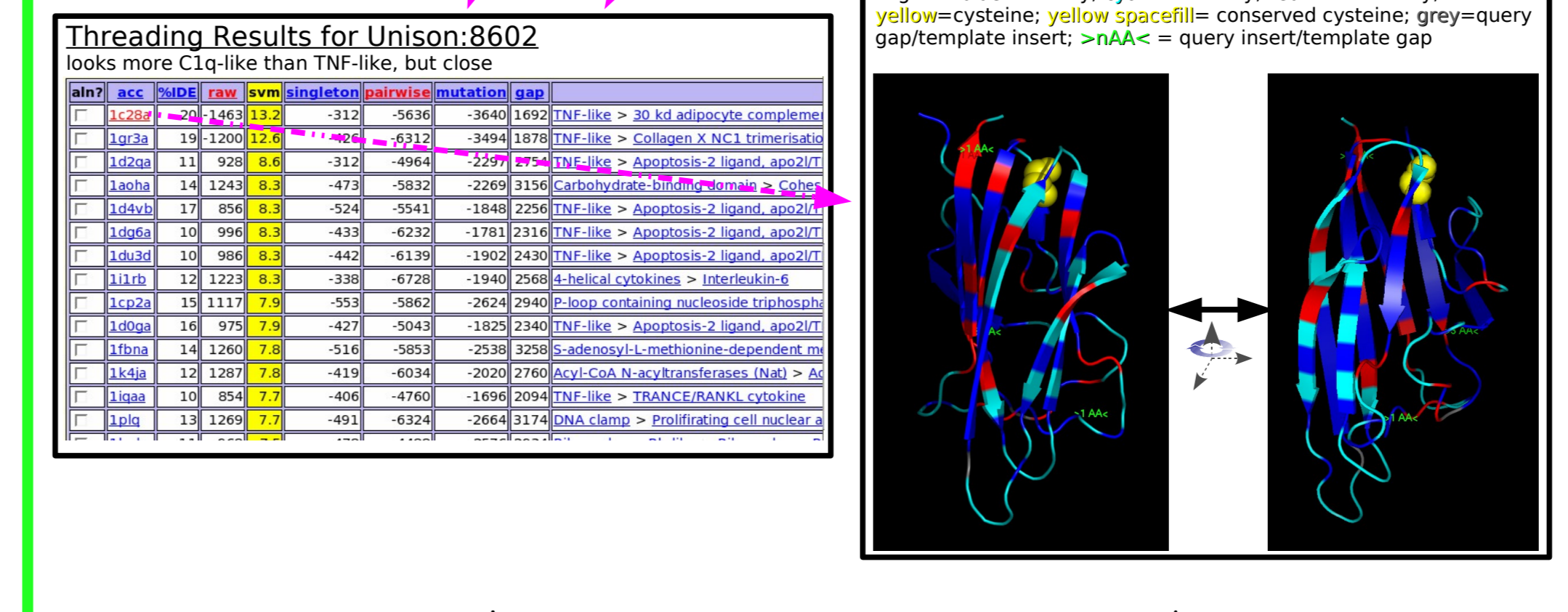
3. Can we use Prospect2 threading instead of Pfam to identify Ig domains?  
Using queries similar to those in (1), we proposed candidates using both Ig and FnIII domains identified by Pfam HMMs or by threading to PDB structures. Using the genomic localization and HomoloGene data in Unison, it was trivial to extend the SQL queries to distinct loci and to sequences which had mouse or rat ITIM orthologs (i.e., orthologs with an Ig domain, TM, and ITIM).



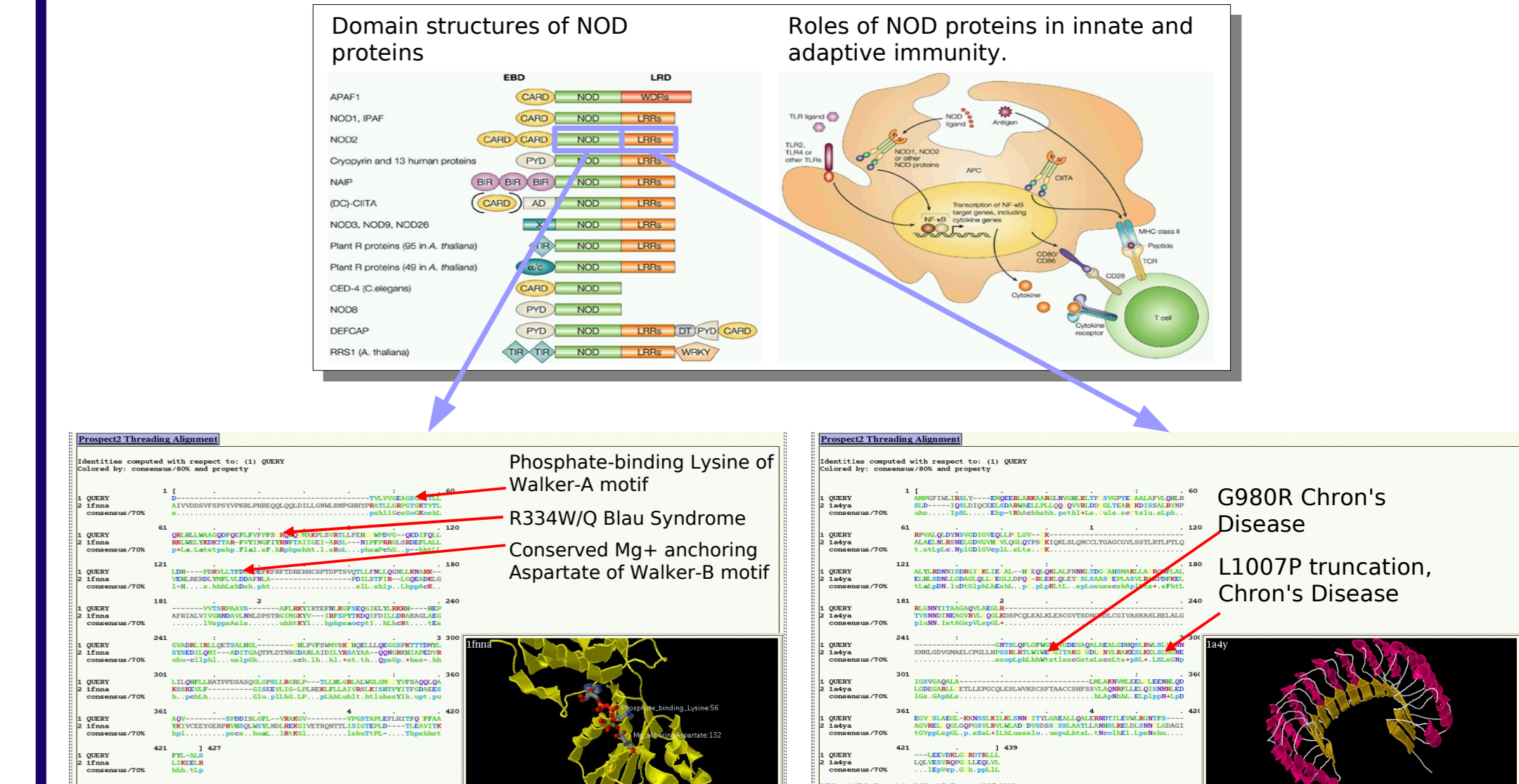
**Mining for Tumor Necrosis Factor Ligand Homologs**  
Tumor Necrosis Factor (TNF) ligands, acting through their cognate TNF receptors, are critical to numerous immunological responses, including B and T cell differentiation, apoptosis, and inflammation. Several "orphan" TNF receptors exist for which the corresponding ligands are unknown. Over the past several years, we have undertaken attempts to identify these unknown ligands from curated protein sequences, six-frame translations of the human genome, and from pathogenic sequences. Because TNF ligands exhibit strong structural conservation and very poor sequence conservation (9-30% identity, avg. 19%), we have concentrated on fold recognition methods.

Integrated search methods. A single Unison page allows users to select and integrate results from HMMs, PSSMs, and Prospect2 threadings to any family of models (TNFs in this case). "Hits" are classified into true positives, false negatives, and "unknown" positives (candidates) by reference to a curated list of known family members.

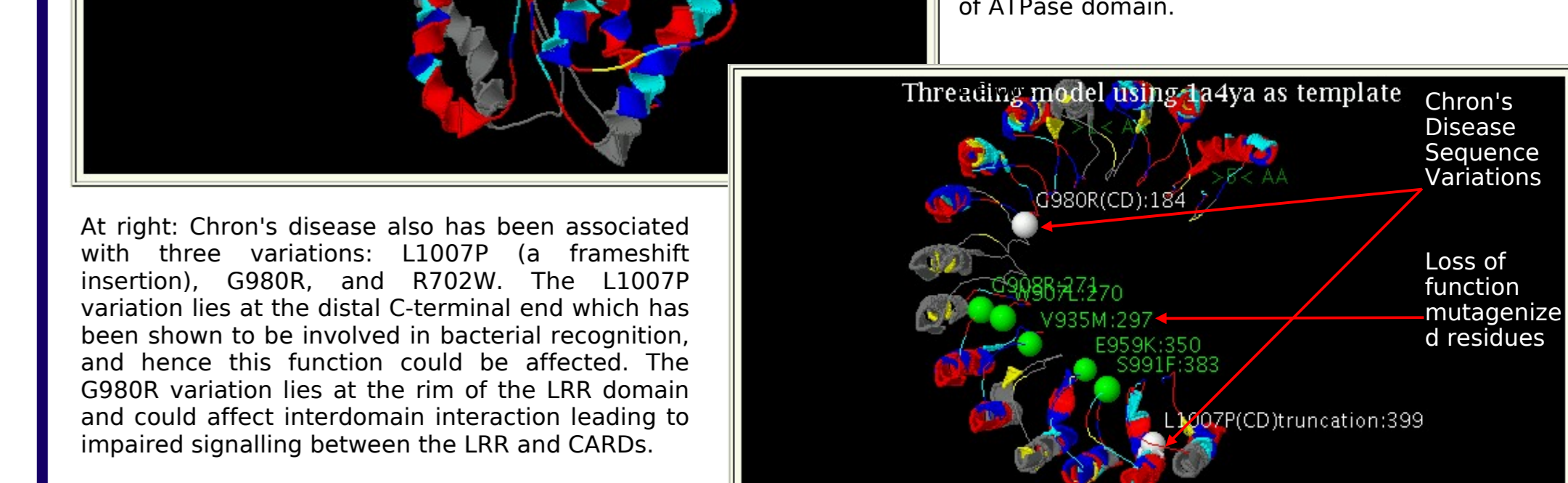


**Modeling NOD2 Polymorphisms**  
Sequence variations in the NOD2 gene cause at least two distinct autoinflammatory diseases: Blau-Syndrome, an autosomal dominant trait, and Chron's Disease, a common inflammatory disease of the intestine. The molecular mechanism of these SNPs is unknown.



Threading-based alignment and homology model of NOD2 NACHT domain. (The structure of Apaf1 was recently solved and agrees well with this model.)

Unison screenshots showing SNPs on homology models. This aspect of Unison is under active development. It is not yet part of publicly available Unison source code (but stay tuned!).



The assessment of NOD2 polymorphisms in this study emphasizes the utility of integrating sequence, precomputed structure predictions, genome variations, and visualization tools within a single environment.

**Acknowledgments**

- Kiran Mukhyala and David Cavanaugh have contributed immensely to Unison.
- Thanks to Hilary Clark and Daryl Baldwin for discussions regarding ITIMs.
- The TNF mining effort was a multi-year collaboration within Genentech and included: Vishva Dixit, Wayne Fairbrother, Sarah Hymowitz, Nobuhiko Kayagaki, Nick Skelton, Minhong Yan, and Zemin Zhang.
- Kiran was responsible for the NOD2 analyses.
- Thanks to Genentech and William Wood for providing a great place to work.

See also our poster in Poster Session 2 on Monday.