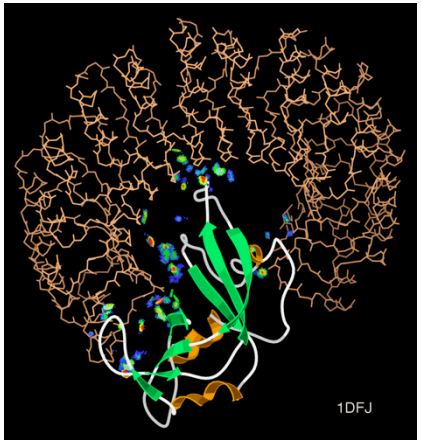
A computational interactome and functional annotation for the human proteome

Garzón, Deng, Murray, Shapira, Petrey, Honig

Protein-Protein Interaction Prediction



Wikipedia

Protein-Protein Interaction Prediction can:

1. Help us understand the mechanisms behind disease
2. Provide more druggable targets

Goal: map all of the interactions in a given set of proteins

Framed as: a learning problem of predicting whether a pair of proteins will interact

The paper computationally builds a partial interactome of the human proteome.

Predicting Protein-Protein Interactions Database

The human body has approximately 20,000 proteins.

This iteration of PrePPI predicts 1.35 million interactions between 17200 proteins in humans.

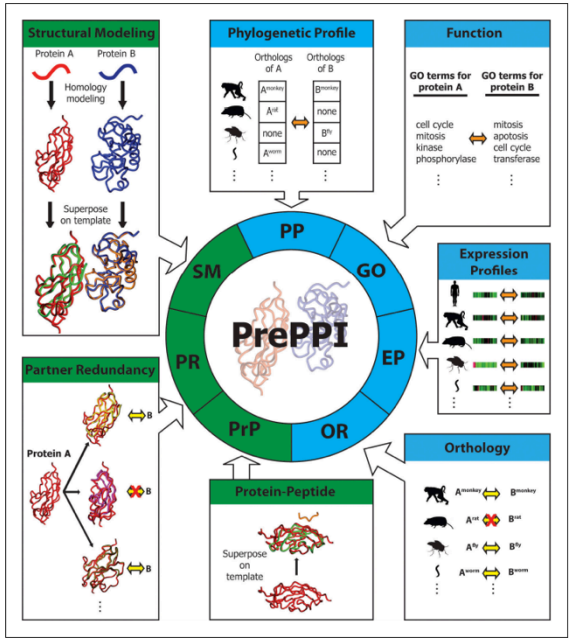
PrePPI, combined with Gene Set Enrichment Analysis (GSEA) allows for the functional annotation of the human proteome.

PrePPI Overview

PrePPI uses both structural and non-structural information to predict interactions without using finely-detailed models or other experimental information.

Information used:

1. Structural Modeling
2. Phylogenetic Profile
3. Gene Ontology
4. Orthology
5. Expression Profile
6. Partner Redundancy
7. Protein Peptide



General Algorithmic Structure

Compute 7 scores based on the given information.

Fit a naive Bayes model with experimental PPIs and their scores.

Training set: Amalgamation of 42,636 yeast PPIs (the yeast HC reference set) reported in at least two publications and a negative (N) set of all pairs in which no interaction is documented in the literature

Test set: Amalgamation of 26,983 human PPIs (the human HC reference set) reported in at least two publications and a negative (N) set of 1,632,716 pairs where each protein belongs to a different cellular component

Structural Modeling

Goal: score an interaction between A and B

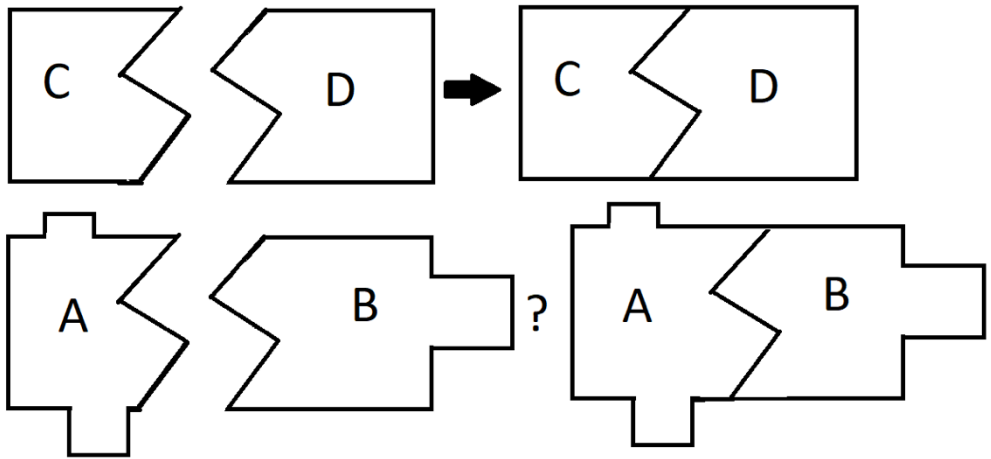
Two structurally similar proteins C and D are picked; the interaction between C and D is known experimentally.

Then the interaction is scored based on the similarities between A and C and B and D and on the interaction between C and D.

For a large number (27726/162833) of the templates, it was unknown whether the protein-protein interfaces was biologically relevant (i.e. not a result of crystallization).

Structural Modeling

A is similar to C and B is similar to D; if C and D interact strongly, then A and B are likely to interact.



Structural Modeling

Given two proteins A and B, find structural representatives MA and MB that correspond to experimental structures or homology models using sequence alignment.

Find all structural neighbors of MA and MB using structural alignment (around 1500 per protein).

Whenever the neighbors form a complex in the PDB, they’re used as an interaction model.

3 values derived from structural modeling: similarity of MA and MB to A and B, number of interacting residue pairs in the interaction model that have analogues in the interaction of MA and MB and the fraction that had analogues.

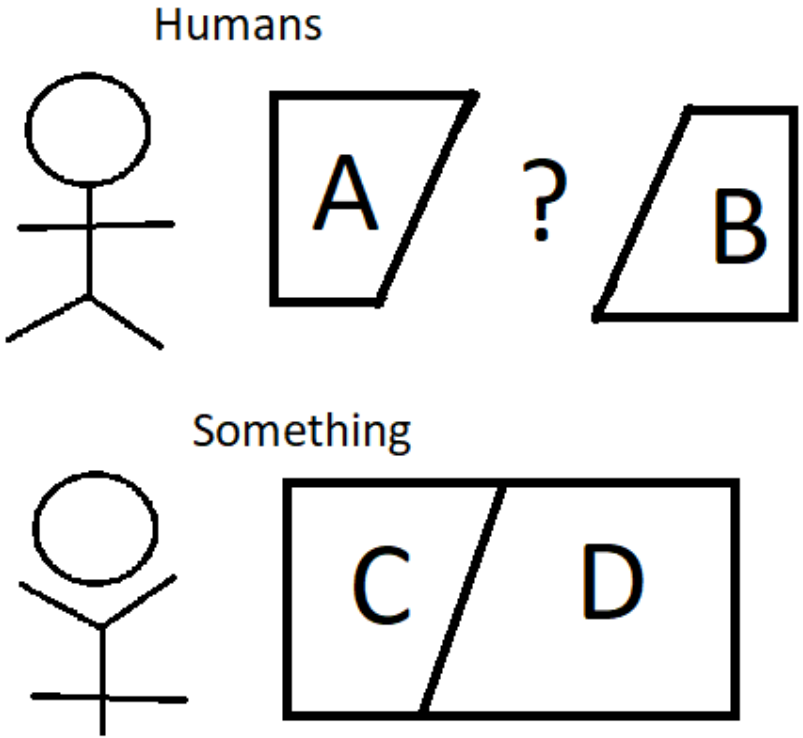
Phylogenetic Profile

Orthologs -- Proteins that have evolved from a common ancestor, but are now in different species; they typically serve similar functions as they did before

The phylogenetic profile score looks at whether orthologs of proteins A and B reside in the same species; this may indicate that they have some dependence on one another.

Originally, phylogenetic analysis was applied in the form of binary vectors on the presence of certain reference proteins in each species, but a tree-based approach has also been proposed.

Phylogenetic Profile



Gene Ontology

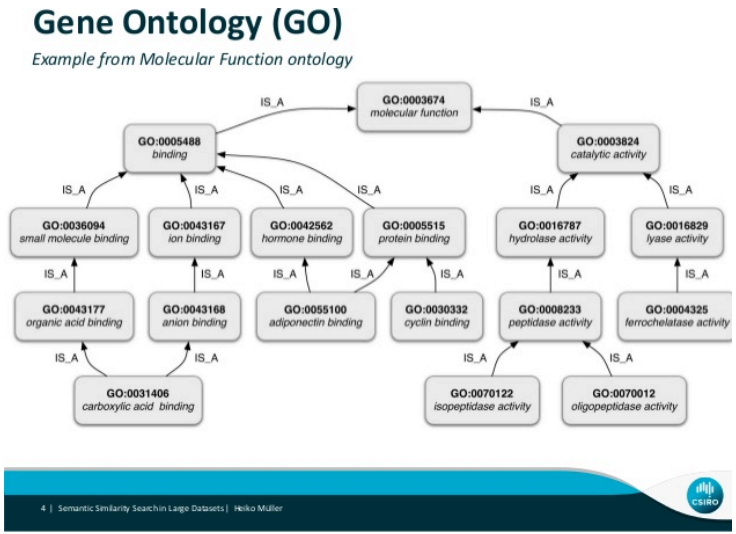
Gene ontology is the process of adding relational annotations between different biological concepts structured in a directed acyclic graph.

3 domains:

1. Cellular component
2. Molecular function
3. Biological process

Metric: distance between the proteins in the graph

Example of an ontological diagram



Orthology

This measures the likelihood that A and B will interact given that their orthologs C and D interact in a different species.

This is similar to the phylogenetic profile, except this considers interaction in a different species.

However, this is still a valuable separate source of data; the correlation coefficient with the phylogenetic profile is a mere 0.003.

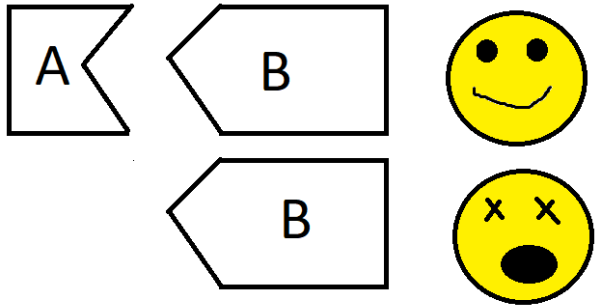
Score is a four-dimensional vector with components 0, 1, or >1 (across 4 databases); there are 81 separate bins.

Expression Profile

The expression profile scores similarities among how the proteins are expressed from the genome.

Improvement over previous version of PrePPI: uses orthologs of A and B in organisms other than humans

Underlying rationale: proteins that must work together are expressed together



Expression Profile

1	0.776197942520391	0.879555282450664	0.653828108793675
0.776197942520391	1	0.828167401726356	0.509637083884152
0.879555282450664	0.828167401726356	1	0.652218961357004
0.653828108793675	0.509637083884152	0.652218961357004	1

Pearson correlation coefficients between each pair of proteins are given from COXPRESdb (Okamura et al.) and ArrayExpress (Kolesnikov et al.).

Let A and B be the two proteins of interest; A_j and B_j are the orthologs in species j.

$$S_{pos}(c_1, c_2, \ldots, c_{n_{pos}}) = 1 - \prod_{j=1}^{n_{pos}} (1 - c_j) \quad S_{neg}(c_1, c_2, \ldots, c_{n_{neg}}) = -\left(1 - \prod_{j=1}^{n_{neg}} (1 - |c_j|)\right)$$

Where n_{pos} is the # of species where the correlation > 0 and n_{neg} is the # where correlation < 0

S_{pos} and S_{neg} are 0 if the correlation is 0.

$$S_{orth} = \begin{cases} S_{pos} & if n_{pos} \geq n_{neg} \\ S_{neg} & if\ n_{pos} < n_{neg} \end{cases}$$

Expression Profile

The COXS (cross-species correlation score) is computed with w = 0.6:

$$COXS(A,B) = \begin{cases} 1-(1-S_{human})*(1-S_{orth}*w) & if\ S_{human} \geq 0\ and\ S_{orth} \geq 0 \\ S_{human} & if\ S_{human} \geq 0\ and\ S_{orth} < 0 \\ S_{orth} & if\ S_{human} < 0\ and\ S_{orth} \geq 0 \\ -(1-(1-|S_{human}|)*(1-|S_{orth}|*w)) & if\ S_{human} < 0\ and\ S_{orth} < 0 \end{cases}$$

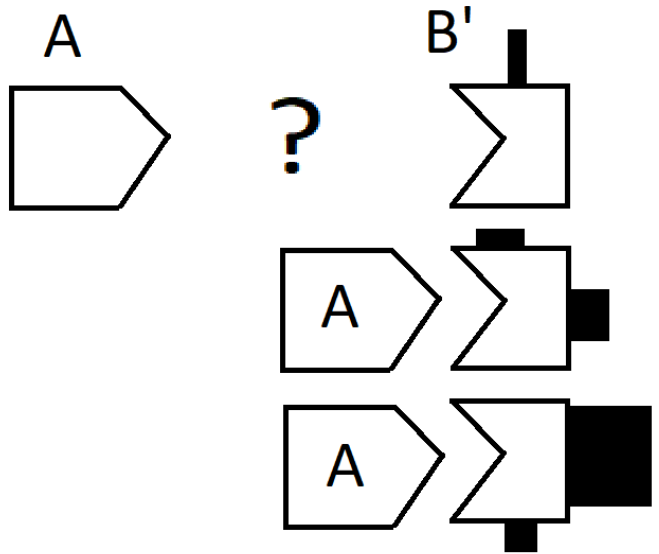
This is binned into 20 equally-sized bins from -1 to 1.

LR values for each bin

[-1.0,-0.9)	0.436383	[0,0.1)	1.175717
[-0.9,-0.8)	0.24982	[0.1,0.2)	1.581839
[-0.8,-0.7)	0.595423	[0.2,0.3)	2.107711
[-0.7,-0.6)	0.888846	[0.3,0.4)	2.771394
[-0.6,-0.5)	0.541017	[0.4,0.5)	3.900837
[-0.5,-0.4)	0.774909	[0.5,0.6)	5.687484
[-0.4,-0.3)	0.668759	[0.6,0.7)	9.468682
[-0.3,-0.2)	0.712574	[0.7,0.8)	18.44262
[-0.2,-0.1)	0.79625	[0.8,0.9)	45.27978
[-0.1,0)	0.691518	[0.9,1]	38.17397

Partner Redundancy

Intuition: If the protein A interacts with proteins B1, B2, B3, B4, …, Bn, which are all structurally similar to B, then the chance that A and B interact gets higher with increasing n.

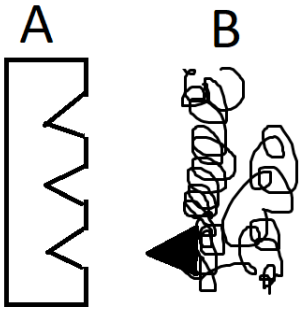


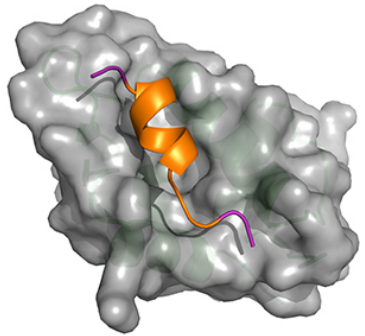
Protein Peptide

Peptide -- short chain of amino acids

This is scored on how likely it is for A and B to interact if A has a certain structure to it and B has a motif (local structure found in unrelated proteins) that is known to interact strongly with that structure.

This is mutually exclusive with the first metric; therefore, the maximum of the two likelihood ratios is used, as opposed to multiplying both together.





Naive Bayes

The paper uses Naive Bayes with binned scores to calculate the likelihood ratios.

Naive Bayes aims to find the probability of observing the *features* given the class.

$$P(\mathbf{x} = (a_1\ a_2\ a_3\ \ldots\ a_m)^T | C_k) = \prod_{i=1}^{m} P(x_i = a_i | C_k)$$

It assumes conditional independence of the features on the class label. In this case, k = 0 is no interaction, and k = 1 is interaction.

Naive Bayes -- Typical Application

Now that $P(x_1, x_2, \ldots, x_m \mid C_k)$ is known, one can calculate $P(C_k \mid x_1, x_2, \ldots, x_m)$ by application of Bayes' theorem, where $P(C_k)$ is the prior of the class.

$$P(C_k|\mathbf{x} = (a_1\ a_2\ a_3\ \ldots\ a_m)^T) = \frac{P(C_k)\prod_{i=1}^{m} P(x_i = a_i|C_k)}{P(\mathbf{x})}$$

To classify a feature vector, **x**, one attempts to maximize the LHS of the above formula by choosing k = 0 or k = 1. $P(\mathbf{x})$ is a constant and can be ignored.

However, the paper actually defaults to k = 0 and decides whether k = 1 that discards the prior $P(C_k)$, since this is hard to calculate with regards to PPIs.

PrePPI and Naive Bayes

In PrePPI, the scores of the seven metrics are distributed into bins (discretized), and then a Naive Bayesian network is used (one parent node, a lot of child nodes). Essentially, the joint conditional probability is just the product of the individual conditional probability.

Likelihood ratio of A conditioned on B is $\frac{P(A|B)}{P(A|B^C)}$

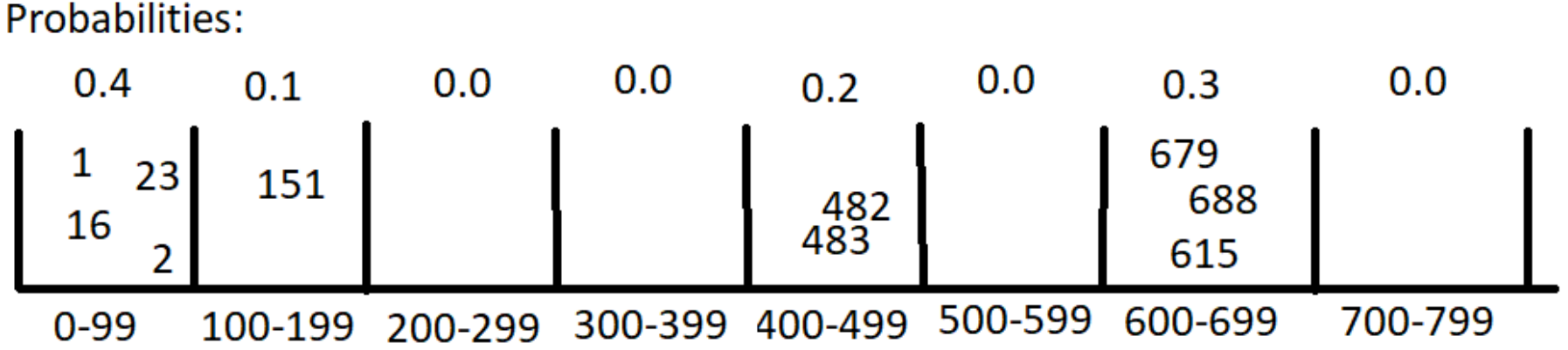
Noting that because conditionally independent probabilities multiply, the likelihood ratios multiply:

$$LR_{\mathrm{PrePPI}} = max(LR^{SM}, LR^{PrP}) \times LR^{PR} \times LR^{GO} \times LR^{PP} \times LR^{OR} \times LR^{EP} \times LR^{PR}$$

$$LR_{\mathrm{PrePPI}} = \frac{P(\mathbf{x} = (a_1\ a_2\ a_3\ \ldots\ a_m)^T|C_1)}{P(\mathbf{x} = (a_1\ a_2\ a_3\ \ldots\ a_m)^T|C_0)}$$

Binning

The score for each of the 7 sources of information is placed into an appropriate bin (discretized) to approximate $P(x_i = a_i \mid C_k)$ in the application of the Naive Bayes algorithm.



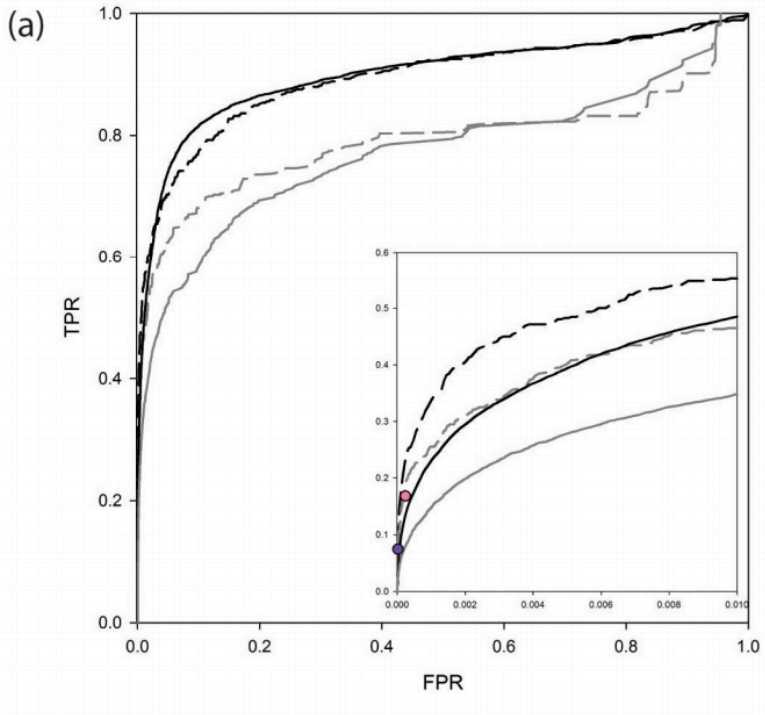
Performance

Evaluated against test set and additional test set (500 in PRS and 700 in RRS) from Rolland et al with receiver operating characteristic as the metric.

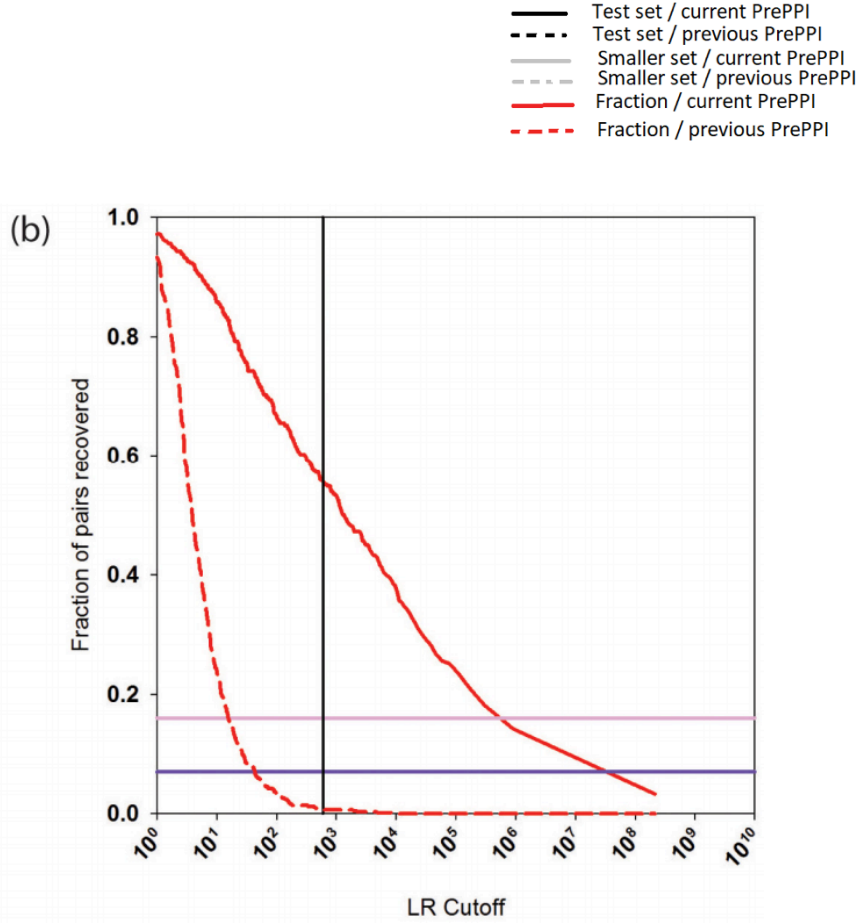
The black curves are for the large test set; the gray ones are for the test set from Roland et al; the red curves are the percent of pairs extracted with a given likelihood ratio threshold.

The dashed curves represent a previous iteration of PrePPI.

Performance



Receiver Operating Characteristic

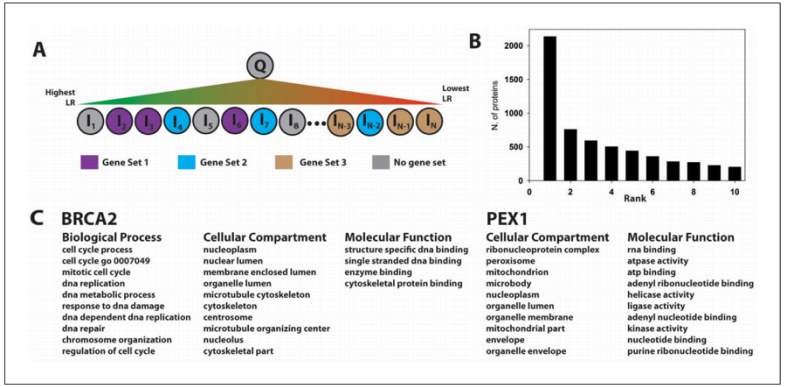


Fraction of recovered pairs vs LR cutoff

Functional Annotation (modified GSEA)

Inference for a given protein Q is accomplished by sorting the human proteome by LR and then looking for a GO annotation among the proteins that strongly interact with Q and whose functions are known.

For 2100 of the proteins, the most enriched gene set is associated with the correct biological process, molecular function, or cellular component; for 5500, the correct GO annotation is found within the top ten gene sets. (Out of 10800)



Other Work

Substantial database overlap with smaller computational databases is validation of the PrePPI database.

However, PrePPI is much larger, possibly containing many more useful interactions.

Overlap of PrePPI with other computational databases

PrePPI	Y2H	BioPlex	PIP	I2D Ophid	HumanNet	String	Comp. All	Hum.Exp.	
1,354,007	972	5364	17,639	67,556	76,905	123,457	212,463	44,864	PrePPI
	13,584	425	140	13,470	804	918	13,584	1777	Y2H
		56,553	918	5689	4361	5549	56,553	4763	BioPlex
			44,148	6253	10,324	703	44,148	5154	PIP
				296,008	56,584	53,178	296,008	160,581	I2D Ophid
					458,518	58,512	458,518	44,047	HumanNet
						311,635	311,635	45,890	String
							1,004,622	162,065	Comp. All
								169,368	Hum. Exp.

Unique interactions in PrePPI

	PrePPI	Y2H	BioPlex	PIP	I2D Ophid	HumanNet	String	Comp.All	Human Exp.
PrePPI	1,354,007	1,353,035	1,348,643	1,336,368	1,286,451	1,277,102	1,230,550	1,141,544	1,309,143
Y2H	12,612	13,584	13,159	13,444	114	12,780	12,666	0	11,807
BioPlex	51,189	56,128	56,553	55,635	50,864	52,192	51,004	0	51,790
PIP	26,509	44,008	43,230	44,148	37,895	33,824	36,245	0	38,994
I2d Ophid	228,452	282,538	290,319	289,755	296,008	239,424	242,830	0	135,427
HumanNet	381,613	457,714	454,157	448,194	401,934	458,518	400,006	0	414,471
String	188,178	310,717	306,086	303,732	258,457	253,123	311,635	0	265,745
Comp. All	792,159	991,038	948,069	960,474	708,614	546,104	692,987	1,004,622	842,557
Human Exp.	124,504	167,591	164,605	164,214	8787	125,321	123,478	7303	169,368

Strengths

Can distinguish paralogs, even though structure is very similar

Recovered ⅔ of CORUM complexes (indirect interaction) at likelihood of 600

Compared to attempted construction of random CORUM complexes; none were recovered

Applications in functional annotation

Combines previous approaches to PPI prediction into one score, and produces a very large database of PPIs

Weaknesses

Interactions in PrePPI are not verified (beyond the overlaps with human experimental data); this means that PrePPI will have a tendency to make too many false positive predictions

Application to functional annotation has a rather low success rate

Overlap between most PPI databases is low, so there isn’t much validation

Use of Pearson correlation, which only measures linear correlation, to say that other variables contributed new data

Weaknesses (cont.)

Use of Naive Bayes and binning means that higher scores on different metrics do not necessarily mean a higher likelihood ratio (although it is very likely if the metrics and training data are properly chosen)

Bayesian Networks

A Bayesian network is a directed acyclic graph (DAG) that represents the conditional dependence of events.

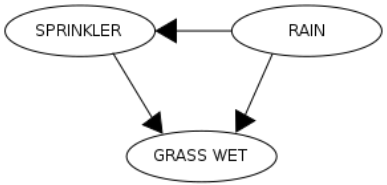
These are different from artificial neural networks in that their structure actually encodes the relationships between the various events.

Each node has a probability distribution conditioned on its parents.

Training is complicated, and inference is NP-hard.

$$
\begin{aligned}
P(F_5|F_1, F_3, F_4) &= \frac{P(F_1, F_3, F_4, F_5)}{P(F_1, F_3, F_4)} \\
&= \frac{\sum_{F_2} P(F_1, F_2, F_3, F_4, F_5)}{\sum_{F_2,F_5} P(F_1, F_2, F_3, F_4, F_5)} \\
&= \frac{\sum_{F_2} P(F_1|F_4)P(F_2)P(F_3|F_5)P(F_4)P(F_5|F_2, F_4)}{\sum_{F_2,F_5} P(F_1|F_4)P(F_2)P(F_3|F_5)P(F_4)P(F_5|F_2, F_4)} \\
&= \frac{P(F_1|F_4)P(F_3|F_5)P(F_4)\sum_{F_2} P(F_2)P(F_5|F_2, F_4)}{P(F_4)P(F_1|F_4)\sum_{F_2,F_5} P(F_3|F_5)P(F_2)P(F_5|F_2, F_4)}.
\end{aligned}
\tag{2}
$$

SPRINKLER		
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

GRASS WET			
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Definitions

Proteome: “the entire set of proteins expressed by a genome, cell, tissue, or organism at a certain time" -- Wikipedia

Interactome: “the whole set of molecular interactions in a particular cell” -- Wikipedia

Functional Annotation: what the proteins and their interactions do to keep things alive

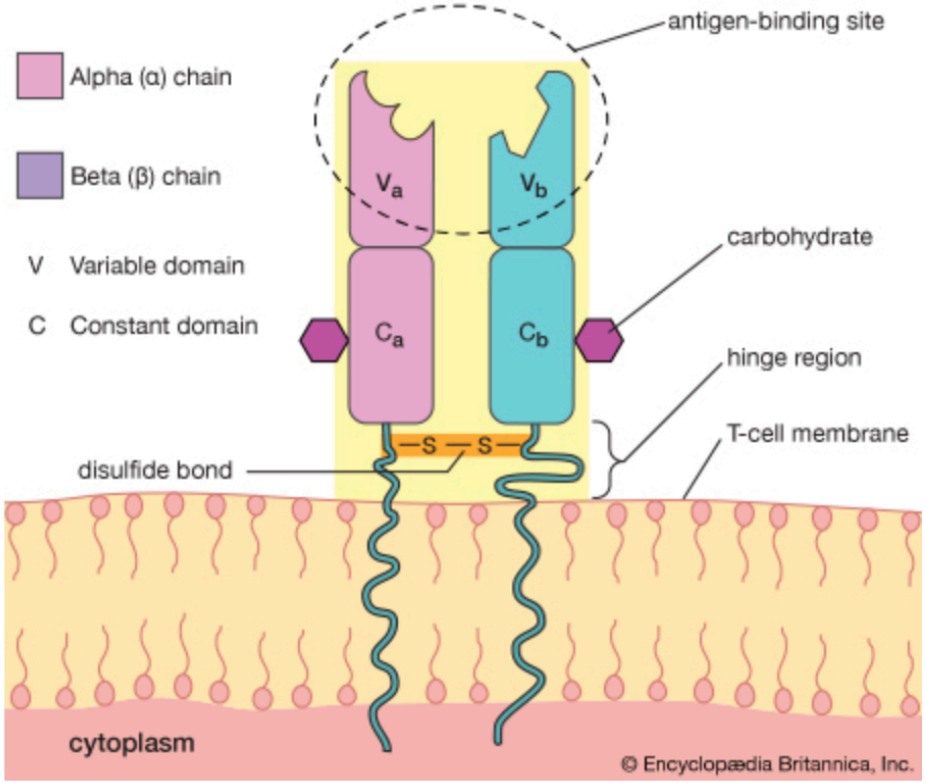
Protein Domain: “a conserved part of a given protein sequence and (tertiary) structure that can evolve, function, and exist independently of the rest of the protein chain” -- Wikipedia

Quantifiable predictive features define specific T cell receptor repertoires

Pradyot Dash, Andrew J. Fiore-Gartland, Tomer Hertz, George C. Wang, et. al

T Cell Receptors

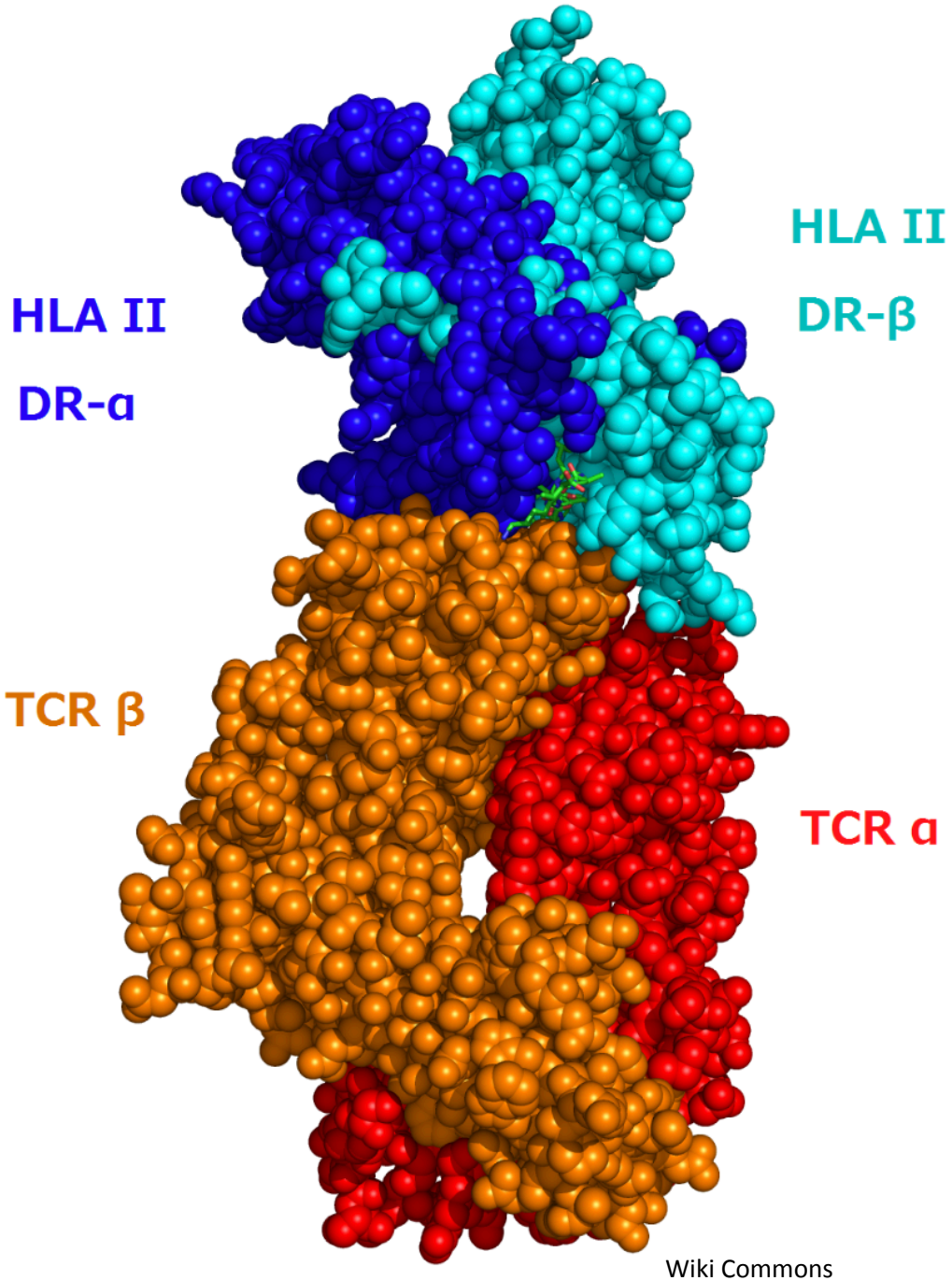
- Mediate recognition of invaders through interactions with peptides from invaders.
- Generated genomic rearrangement => tremendous diversity
  - Can generate 10^15 – 10^62 possible receptors
- Each person can have 100 million different receptors….but….
- T cells that recognize same invader often share conserved features!



https://www.britannica.com/science/T-cell-antigen-receptor

TCR Structure

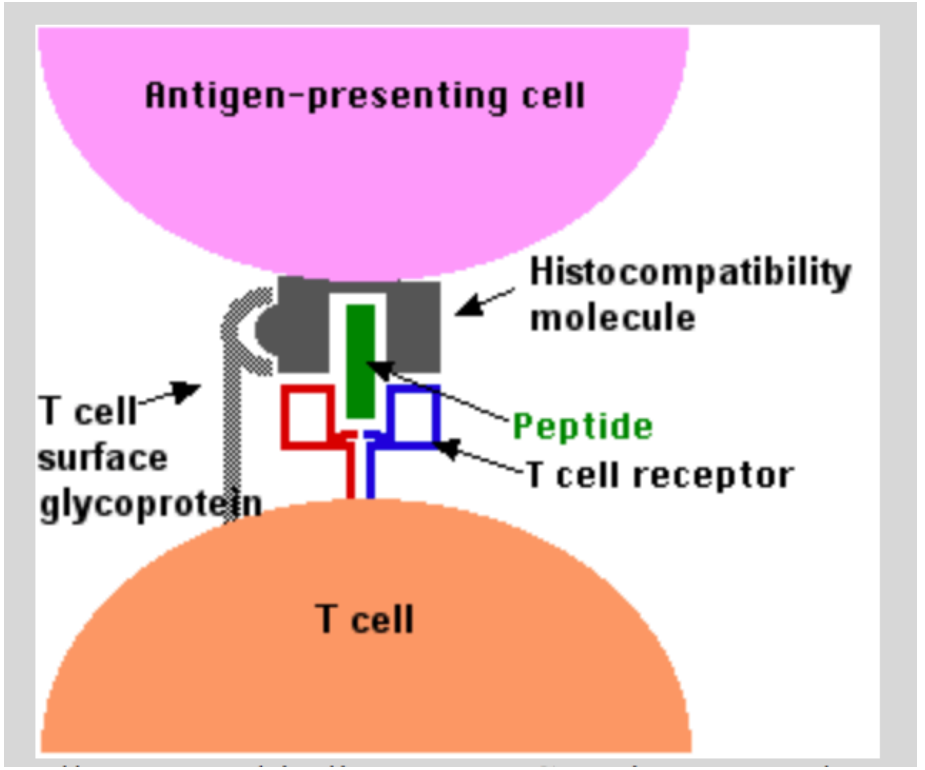
- Membrane anchored heterodimeric protein
- Consists of highly variable alpha and beta chains
- Chain is composed of two extracellular domains
  - Variable: binds to peptide/MHC
    - 3 hypervariable domains: CDRs (mostly CDR3) → mainly responsible for recognizing the antigen
  - Constant



Wiki Commons

Epitope

- TCRs bind to are small peptide fragments called epitopes on antigens (virus particles)
- Displayed by Major Histocompatibility Complex (MHC) on surface of cells
  - MHC = cell surface proteins that bind antigens from pathogens and display for recognition by T-cells;
- Same invader can produce multiple epitopes
- Each epitope is targeted by TCR that are different yet specific



https://www.ebi.ac.uk/interpro/potm/2005_3/Page1.htm

Goal of paper

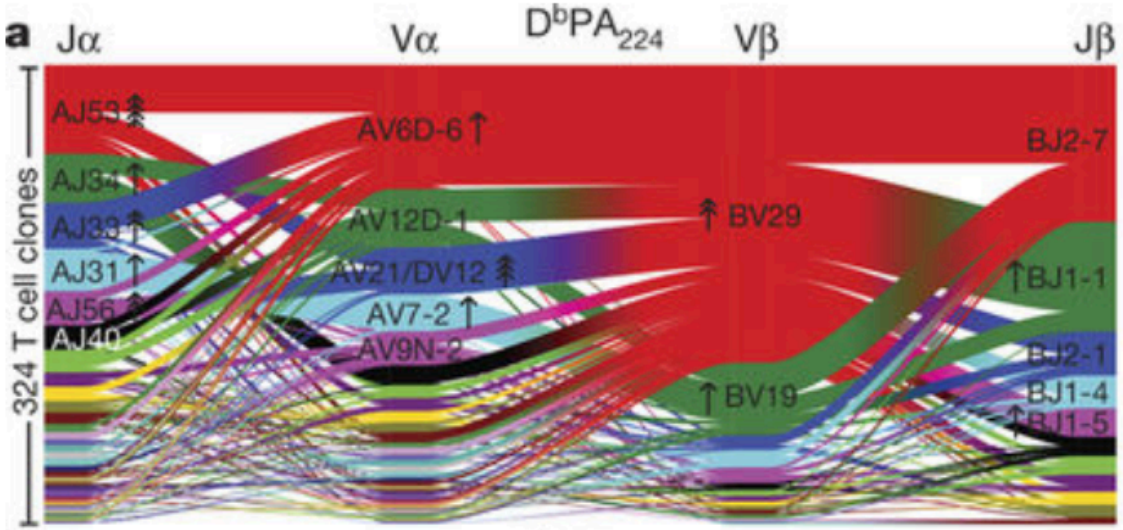
- Goes beyond just establishing that T-cells are diverse: rather, the paper describes how an entire repertoire (group) of T-cell receptors respond to specific epitopes and describes tools to analyze and classify these repertoires.
- Identify underlying features of epitope specific repertoires – key conserved residues driving essential elements of TCR recognition
- Grouping related receptors and selecting representative members
- Analyzed 4635 T cell receptors
  - 10 different epitope specific repertories

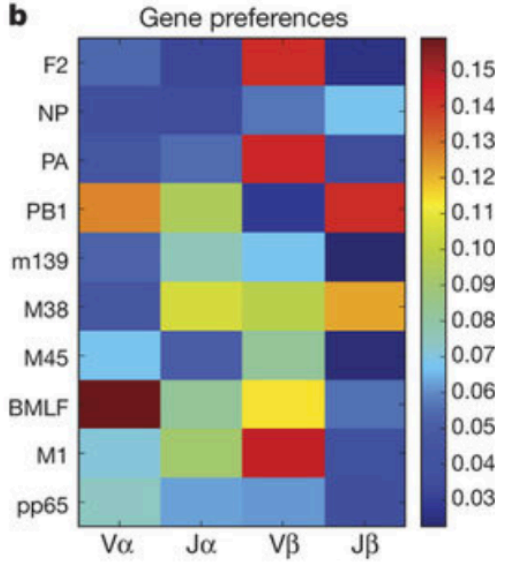
Why does this matter?

- Potential diagnostic tool
- Designing receptors to treat viruses and/or cancer
  - Use immune system as a vehicle to target new viruses or mutations

What prompted this study?

- Each epitope specific response is characterized by an overrepresentation of individual genes
- TCRs that respond to the same epitopes share some features
- Hypothesized that patterns highly unlikely given native distributions must have been selected for and more likely to contribute to specificity





Quantifiable predictive features define epitope-specific T cell receptor repertoires

Metrics to analyze key shared features

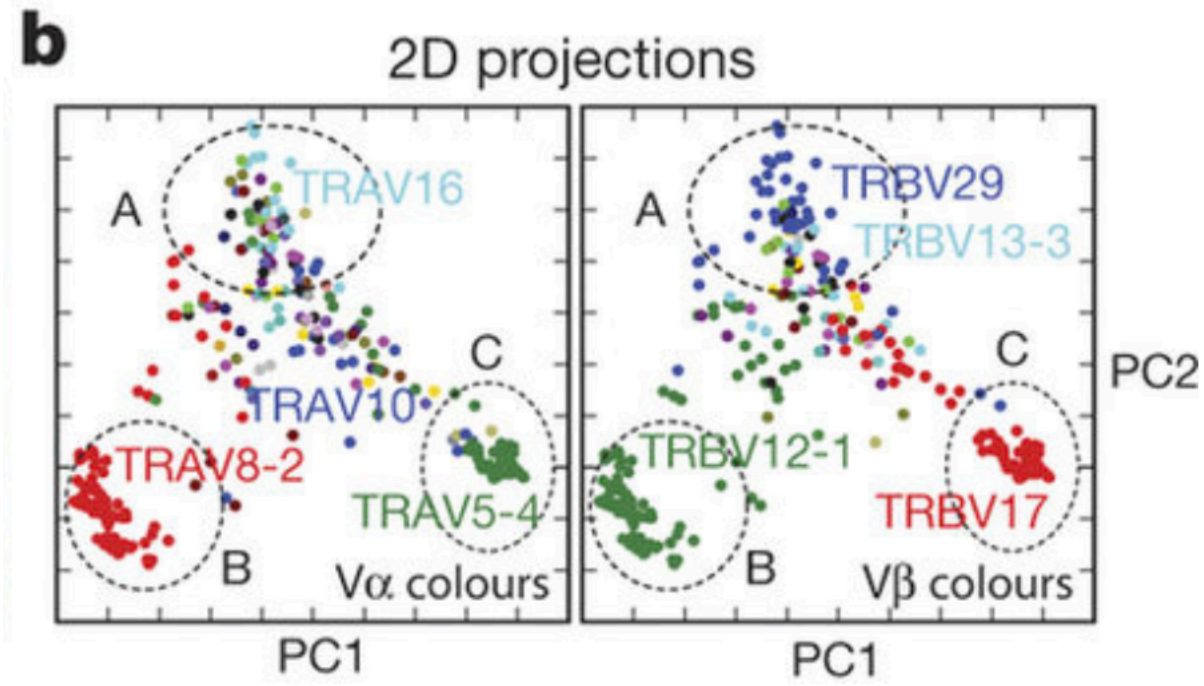
- TCRDist
- TCRDiv
- Nearest Neighbors

TCRdist

- Quantitative measure of similarity between TCR guided by structural information on binding to peptide/MHC complex
  - Similarity between binding regions of different receptors
- Calculated the similarity and differences of key features of T cell receptors
  - Compares amino acid sequences in binding regions
- Matrix of distances between receptors

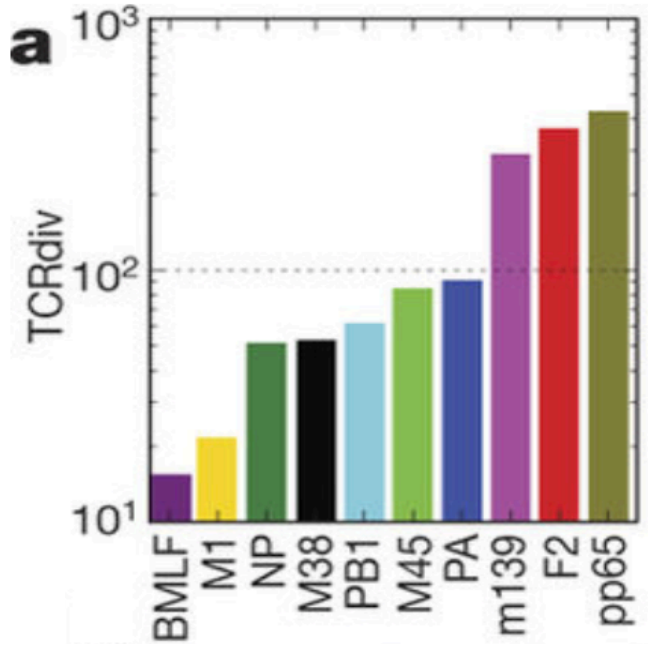
TCRdist

- Visualization of each repertoire by mapping high dimensional TCR landscape into two dimensions: Kernal PCA
  - Look at subregions within each repertoire with similar receptors
- Identified T cell receptors that recognized the same epitope
  - Grouping receptors to see underlying common features



TCRdiv to measure diversity

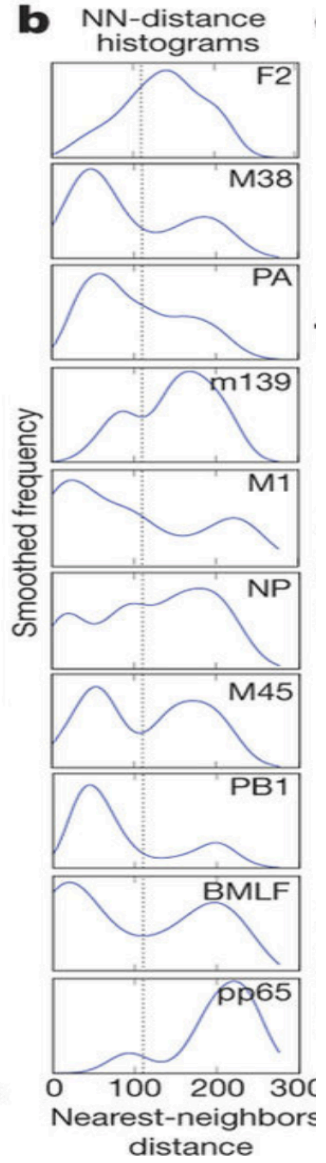
- Generalizes Simpson's Diversity Index (probability of drawing the same class of item in two independent samples) using a distribution of TCRdist values
- Echoed trends seen in TCRdist
- Each repertoire is composed of one more more groups of clustered receptors sharing similar features together with a more diverse population of diverged receptors.



Quantifiable predictive features define epitope-specific T cell receptor repertoires

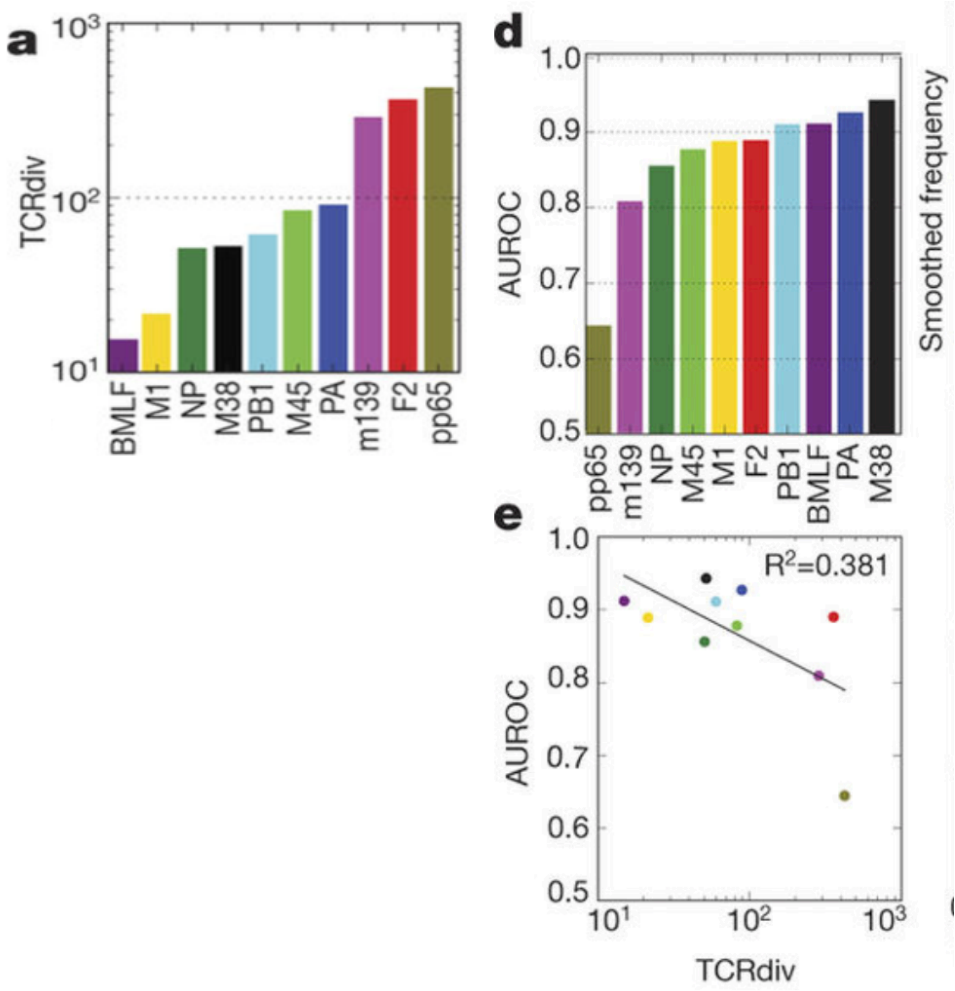
Nearest Neighbors Classifier

- Nearest Neighbor score: density of receptors surrounding each individual receptor
  - a small nearest-neighbours distance → many other nearby receptors → greater local sampling density.
- Majority of epitopes exhibited an approximately bimodal distribution
  - One peak: densely sampled main clusters
  - Second peak: outlier receptors
- Designed TCR classifier that assigns a given receptor to the repertoire with lowest nearest-neighbor distance



Relationship between TCRDiv and classification success

- Measured sensitivity and specificity of classifier for identifying epitope specific receptors among a pool of randomly generated background receptors
- Measured area under false positive and true positive curve (AUROC)
- Most diverse receptors being more difficult to reliably discriminate from the background



Quantifiable predictive features define epitope-specific T cell receptor repertoires

Performance of Nearest Neighbors as Classifier

- Multi class discrimination problem: attempted to assign receptors to correct epitope.
  - Correctly assigned 81% of human T cells and 78% of mouse T cells to one of 10 different viral epitopes
- Tested on three flu infected mice
  - Predicted accurately amongst four flu epitopes
    - Greater than 0.90 AUROC score for three; least accuracy 0.72 for F2.
  - Can classify novel antigen specific T-cell receptors
    - 85% of the receptors correctly classified in this validation experiment were not previously observed

Strengths?

Strengths

- Innovative analytical measurements
  - TCRDist
- Validation
  - Testing for specificity and sensitivity
  - Mouse model
- Paired alpha-beta sequencing

Limitations?

Limitations

- Application of this to other epitopes – solely tested on viruses?
  - B-cell antibodies
- Not just computational validation: in vivo testing of generated receptors
- Effect of MHC allele on classifier performance – all the epitopes testing came from same MHC allele (humans have different MHC alleles)
- Current accuracy of epitope identification alone
- Only used ten different epitopes – accuracy with more?

Next Steps

- Looking at incurable viruses and conducting a similar analysis as an assessment of immune response
- In vivo experimentation: generate novel TCR with most common CDR3 sequences
- Looking at a similar phenomenon in antibodies
  - Applying to invaders more than just viruses