



Center for

*Computational
Biology (CCB)*

**Computational Atlases for
Bioinformatics and Genomics**

Christopher Lee

Computational Atlases: the core concept of CCB

- An atlas is an *alignment* of data maps from different domains. It enables querying of relations from multiple domains to construct “the big picture”
- Integrates huge amounts of disconnected information to discover the patterns that represent their internal logic.



Atlases Connect Data Domains

CT

PET

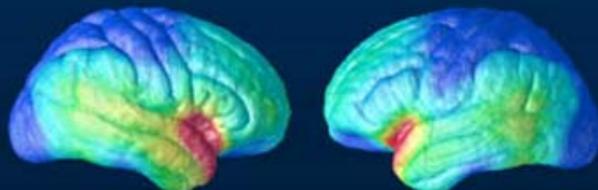
MRI

fMRI

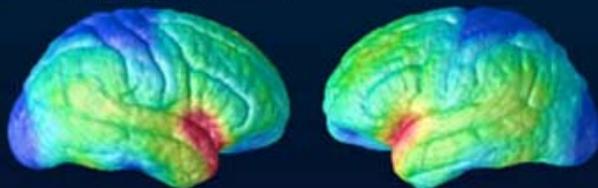


Mean cortical thickness

a HIV (N=27)

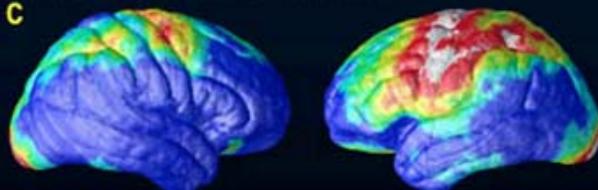


b Healthy controls (N=14)



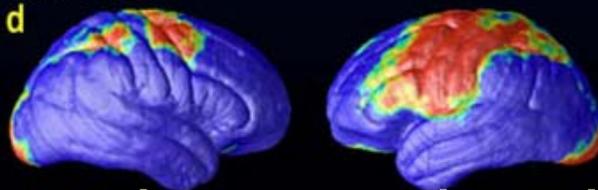
Cortical thickness deficit in HIV

c



Significance

d



Cryo

OIS

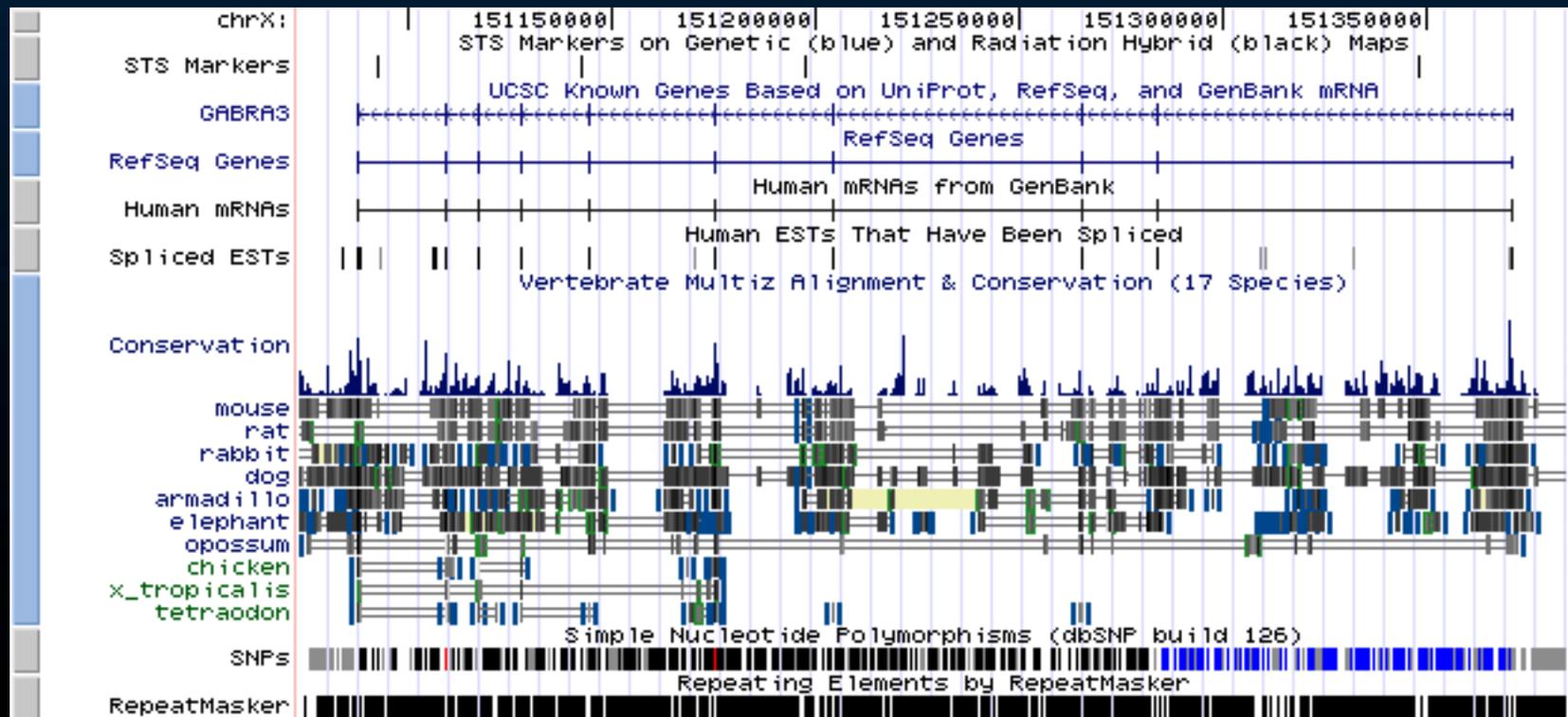


Atlases key for turning raw data into Discovery!



Bioinformatics Needs Atlases Too

- Many separate types of data (DNA, protein, expression, genetics, etc).
- Atlases that map different types of data together have great value.



Queryable Bioinformatics Atlases: a new scale of discovery

- Visualization and browsing is not enough: discovery from “small” data.
- Important to mine patterns from the big picture: discovery from BIG data.
- Not easy to query current databases in a fully integrated way.
- CCB Bioinformatics focuses on this problem.



CCB Bioinformatics Atlases

Tools

- Pygr Graph DB
- NLMSA align't DB
- BLASTgres
- ASAP DB
- 3D Phylogeny
- snpindex
- HIV selection DB

Research

- Alternative splicing
- Comparative genomics
- Atlases of evolution
- HIV drug resistance evolution

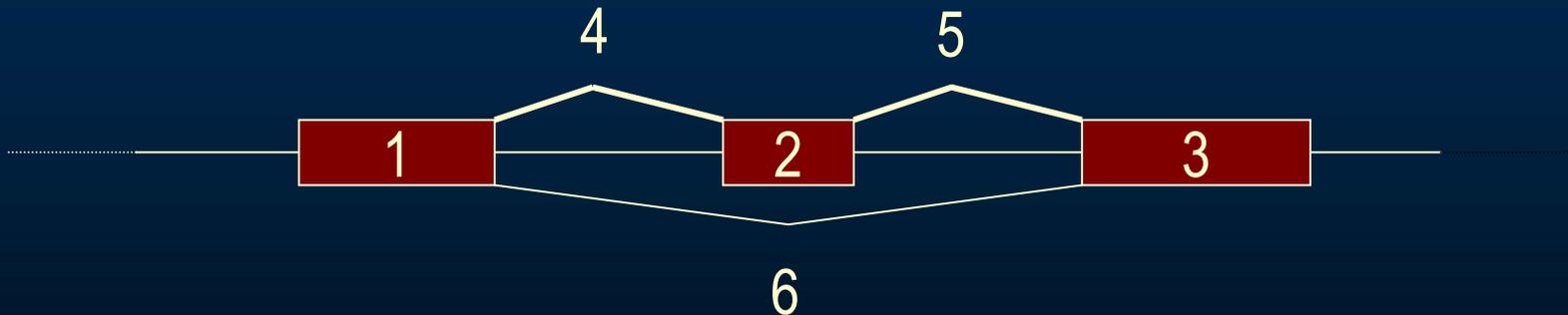


Pygr: A Graph Database Model for Queryable Bioinformatics Atlases

- Relational databases (SQL) have a rigid tabular structure that is often a poor fit to complex biology data.
- Graph databases provide a simpler, more general model: objects (nodes) connected by relations (edges).
- Database is a graph; query is a graph; result is a graph.



Alternative Splicing Example: SQL

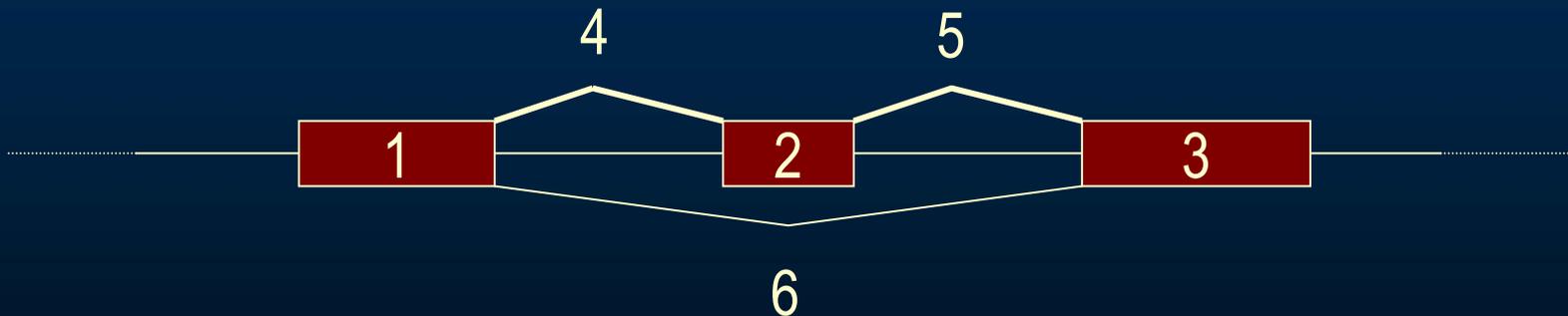


In SQL, a simple exon-skip query requires a 6-way JOIN (groan!)

```
SELECT * FROM exons t1, exons t2, exons t3, splices t4, splices t5,
splices t6 WHERE t1.cluster_id=t4.cluster_id AND
t1.gen_end=t4.gen_start AND t4.cluster_id=t2.cluster_id AND
t4.gen_end=t2.gen_start AND t2.cluster_id=t5.cluster_id AND
t2.gen_end=t5.gen_start AND t5.cluster_id=t3.cluster_id AND
t5.gen_end=t3.gen_start AND t1.cluster_id=t6.cluster_id AND
t1.gen_end=t6.gen_start AND t6.cluster_id=t3.cluster_id AND
t6.gen_end=t3.gen_start;
```



Alternative Splicing Example: Pygr

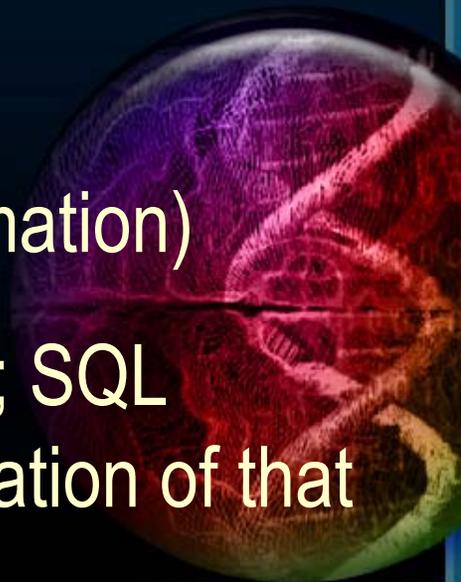


In Pygr, the query graph is just:

$\{1: \{2:\text{None}, 3:\text{None}\}, 2: \{3:\text{None}\}\}$

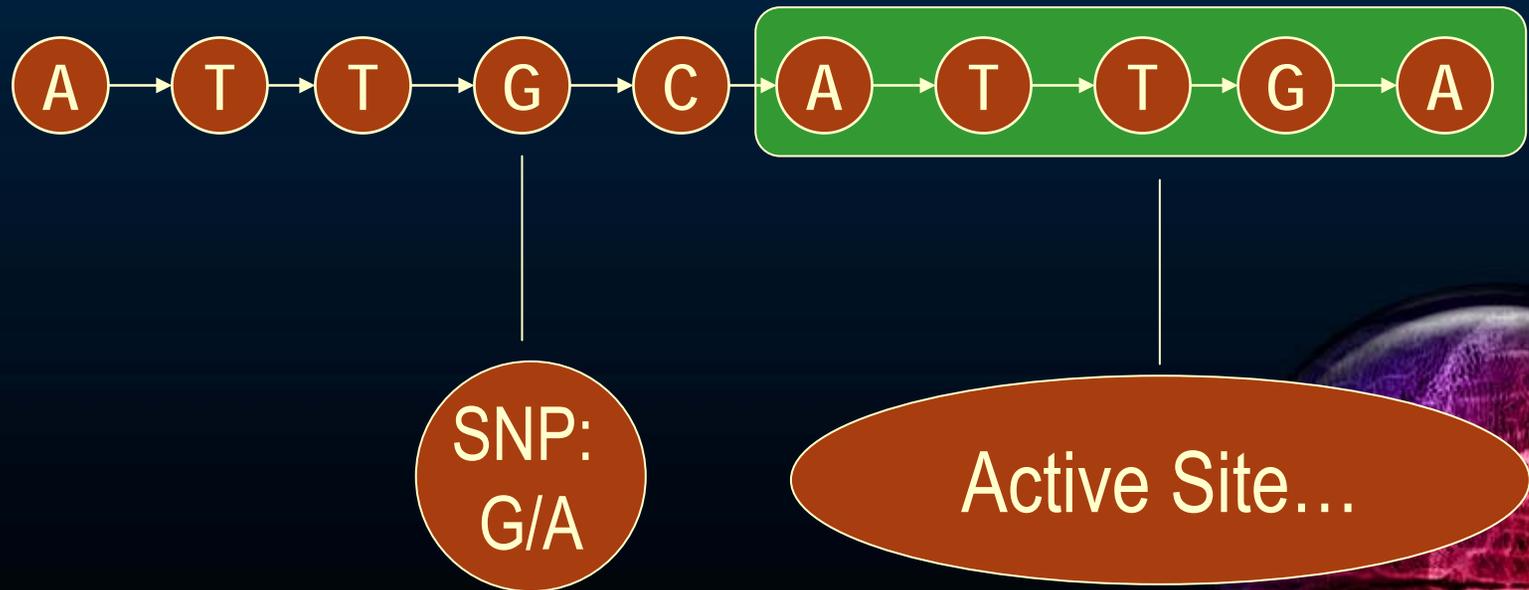
(None means no special edge information)

Simpler because the data *really* is a graph; SQL schema is just a (not very good) representation of that



Hypergraphs are a General Model for Bioinformatics

Sequence Annotation:

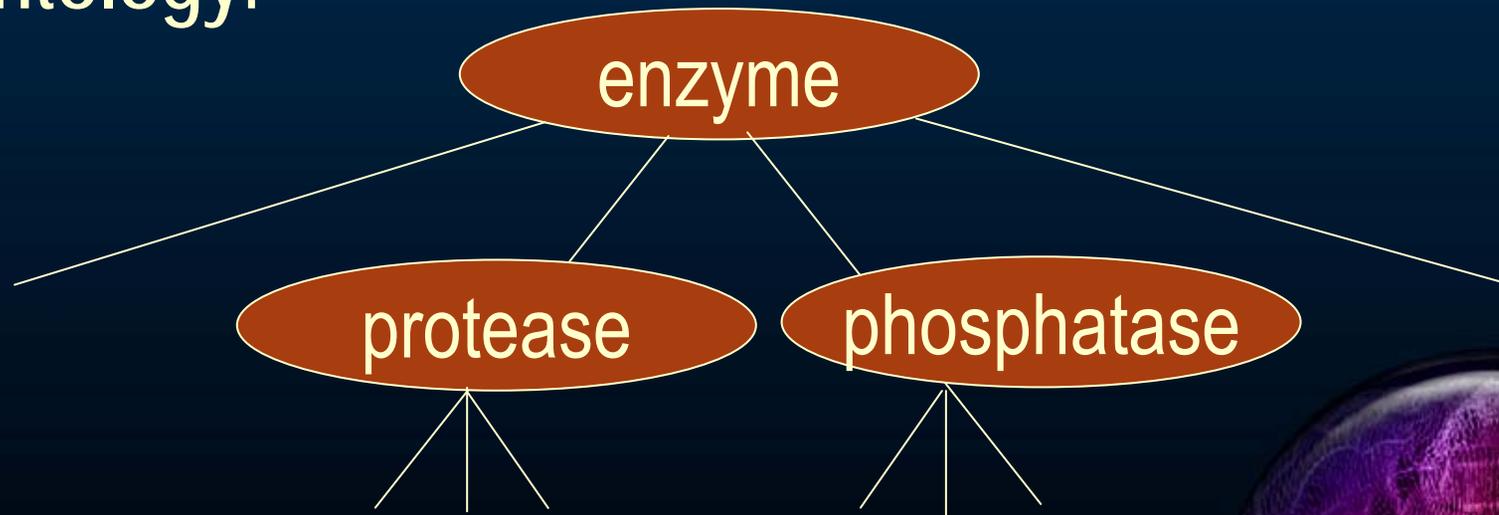


Nodes: sequence letters, annotations

Edges: links between sequence and annotations

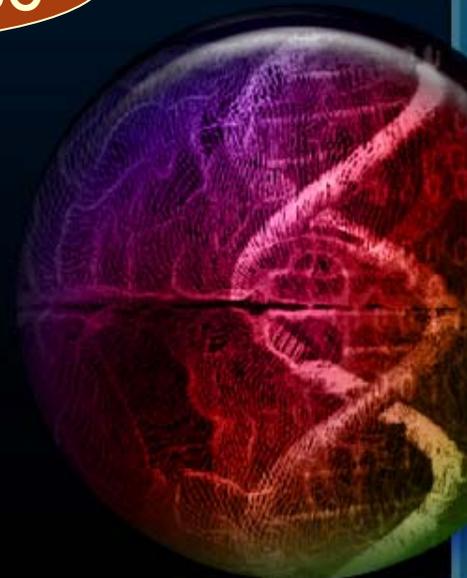
Hypergraphs are a General Model for Bioinformatics

Ontology:



Nodes: terms

Edges: IS-A, HAS-A relations



Pygr works with any data in existing relational databases

Database Tables:

Exons

Splices



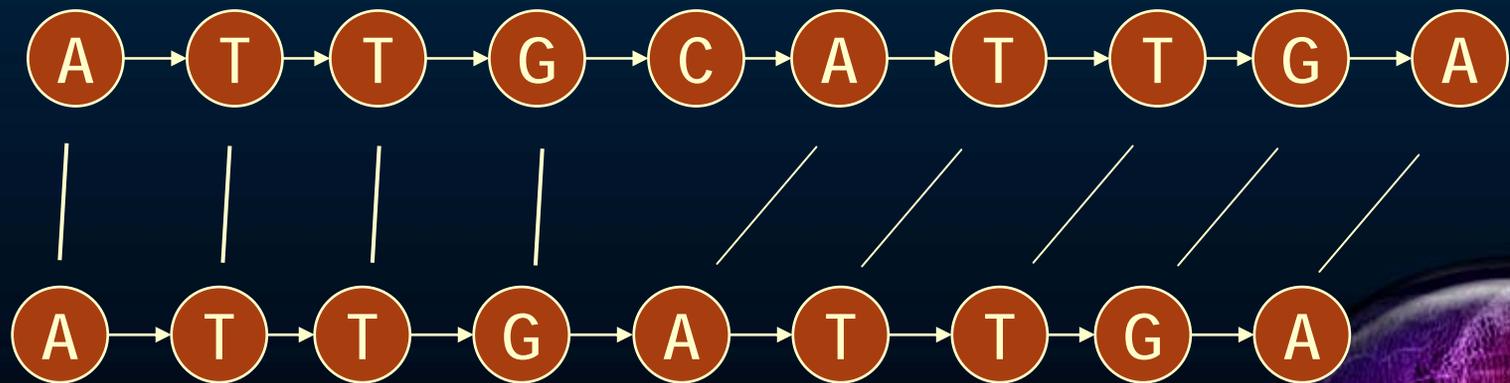
Nodes: rows

Edges: foreign-key relations



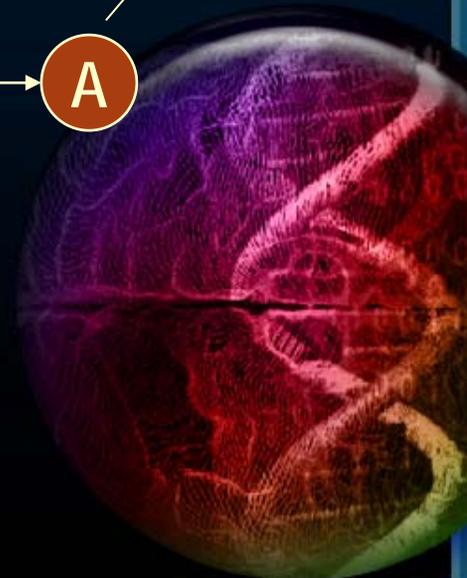
Hypergraphs are a General Model for Bioinformatics

Sequence Alignment:



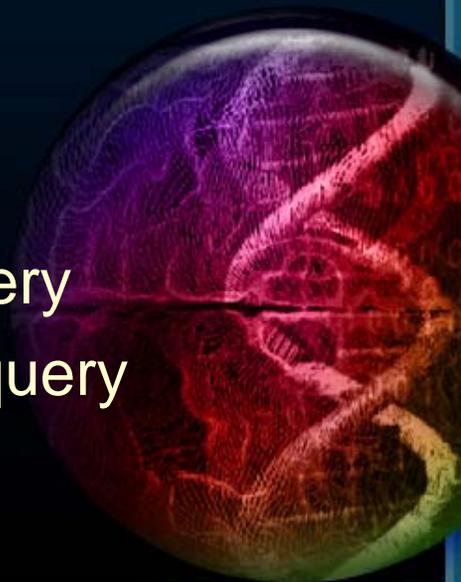
Nodes: sequence letters

Edges: alignment between letters



Graph Database Indexing enables “impossible” queries

- Since all data types are just graphs, a query can traverse different types trivially (hard / impractical in SQL).
- *Edge indexing* accelerates complex relation queries compared with SQL:
 - SQL: >1 hour for exon skip query
 - Pygr: 15 sec. for exon skip query
 - SQL: 30 sec. per genome alignment query
 - Pygr: 0.2 msec per genome alignment query



Example Pygr Applications

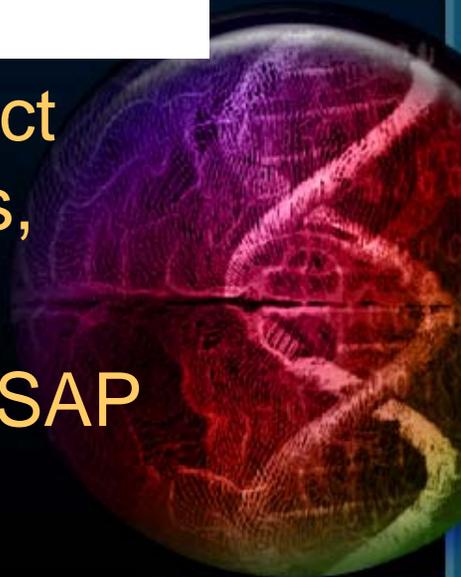
- Analysis of Alternative Splicing
- Analysis of protein domain interaction networks (D. Eisenberg, UCLA)
- Pygr: <http://www.ccb.ucla.edu>
- Python module documentation <http://www.bioinformatics.ucla.edu/pygr>



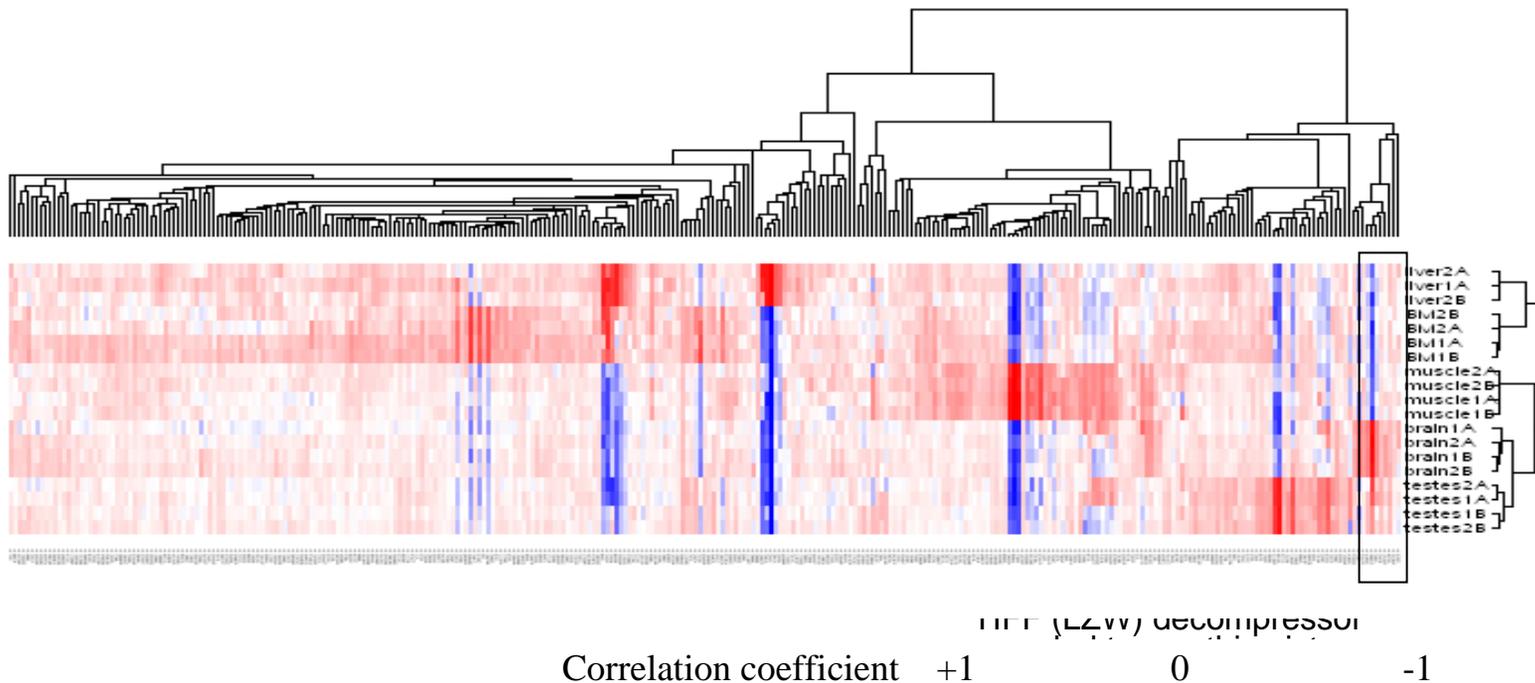
Atlases of Alternative Splicing

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

- Alternative Splicing Annotation Project (ASAP) integrates genomes, mRNAs, ESTs, protein isoforms.
- <http://www.bioinformatics.ucla.edu/ASAP>



Automatic clustering of genes and samples by alternative splicing



Clustered both by samples, and by genes. The clustering revealed groupings of tissue specific alternative spliced genes. For example, a **brain-specific** alternative splicing group is shown at the far right (boxed). This software has also revealed tumor-specific splicing markers.

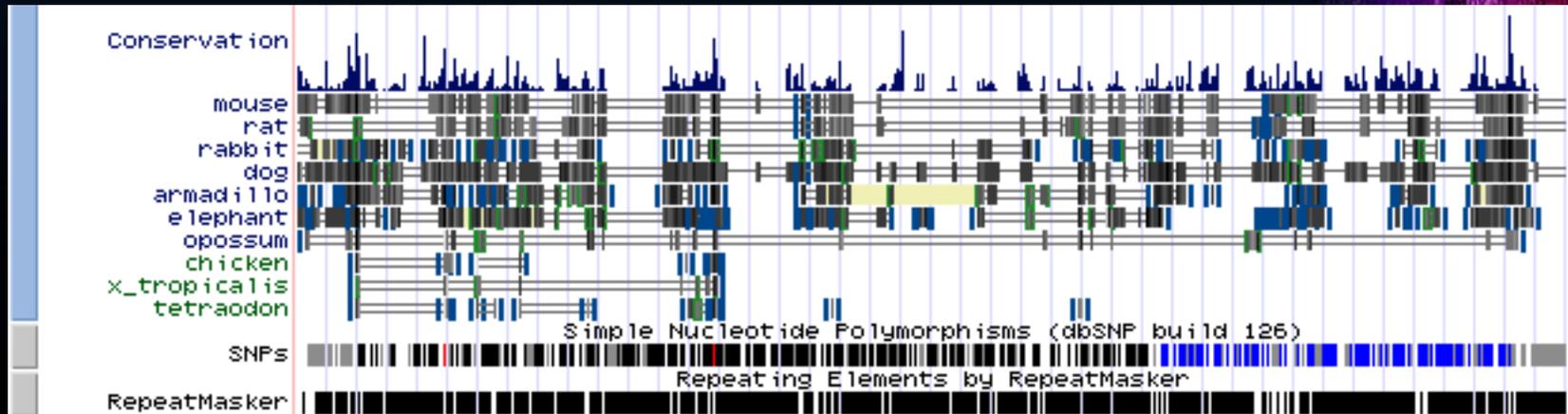
Atlases of Alternative Splicing

- Genome-wide atlas: 90,000 AS events detected for human
- Atlases for 15 animal species
- Tissue-specific AS atlases from EST, microarray data (Blencowe, U. Toronto), especially brain tissues.
- Next: connecting to brain maps via Allen Brain Atlas (expression arrays)

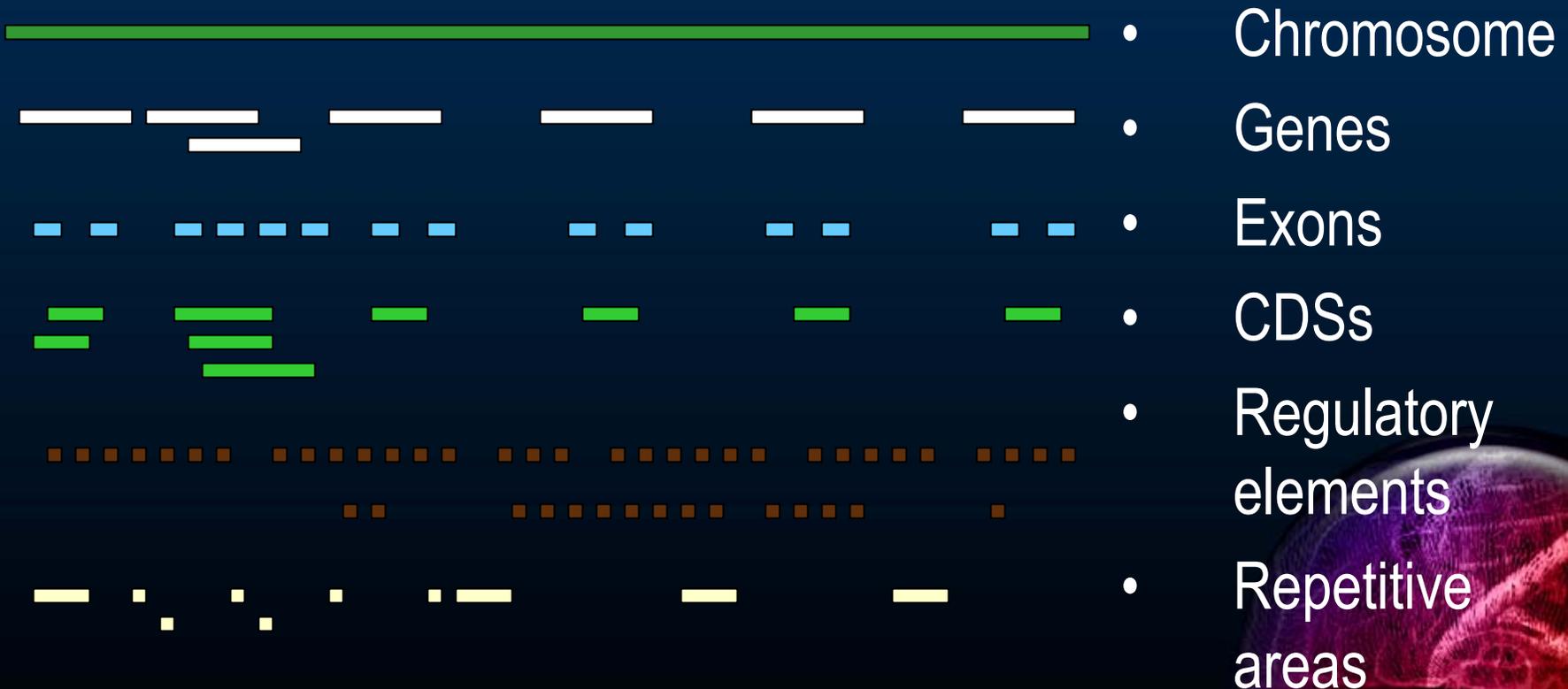


Queryable Atlases of Genome Evolution: NLMSA

- Comparative genomics (genome evolution & conservation) is rapidly becoming important tool for studies of gene regulation, function, disease. Need: query evolution as a *database*

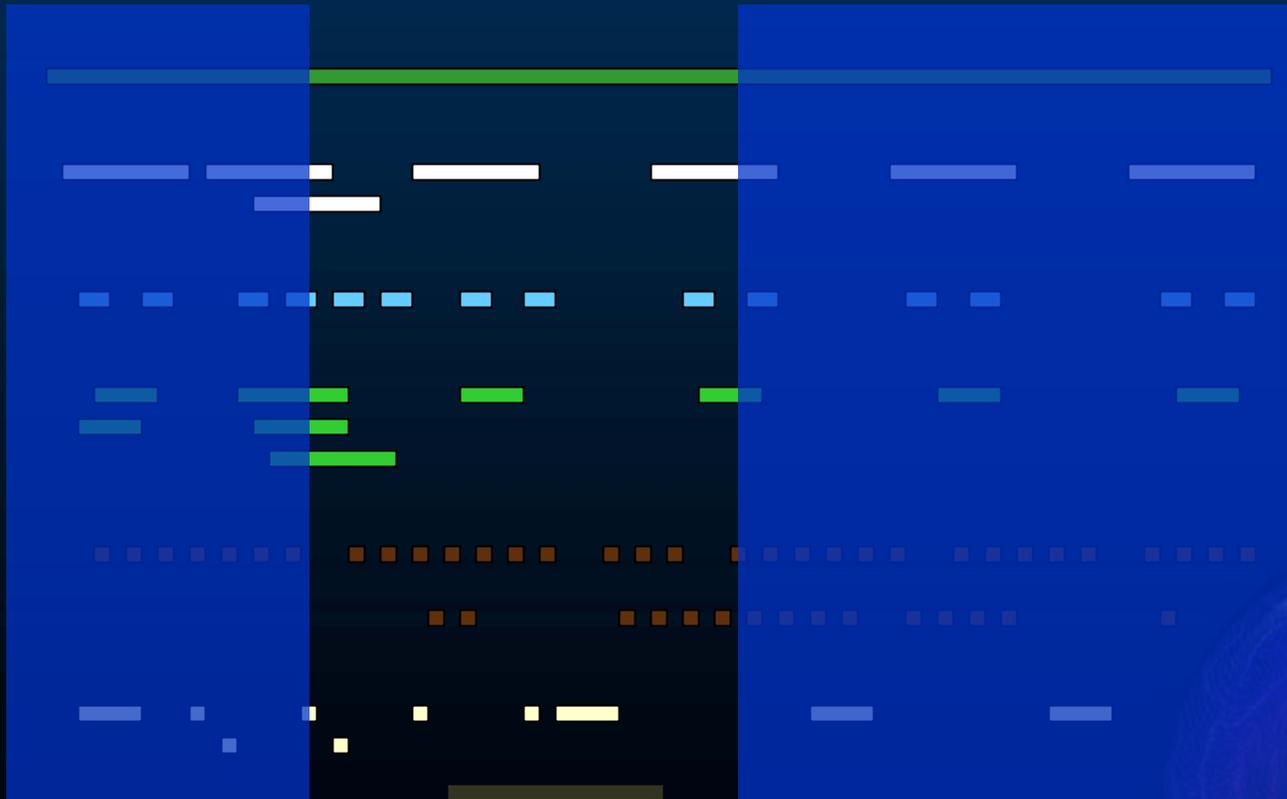


Defining fundamental objects in genome-wide sequence analysis



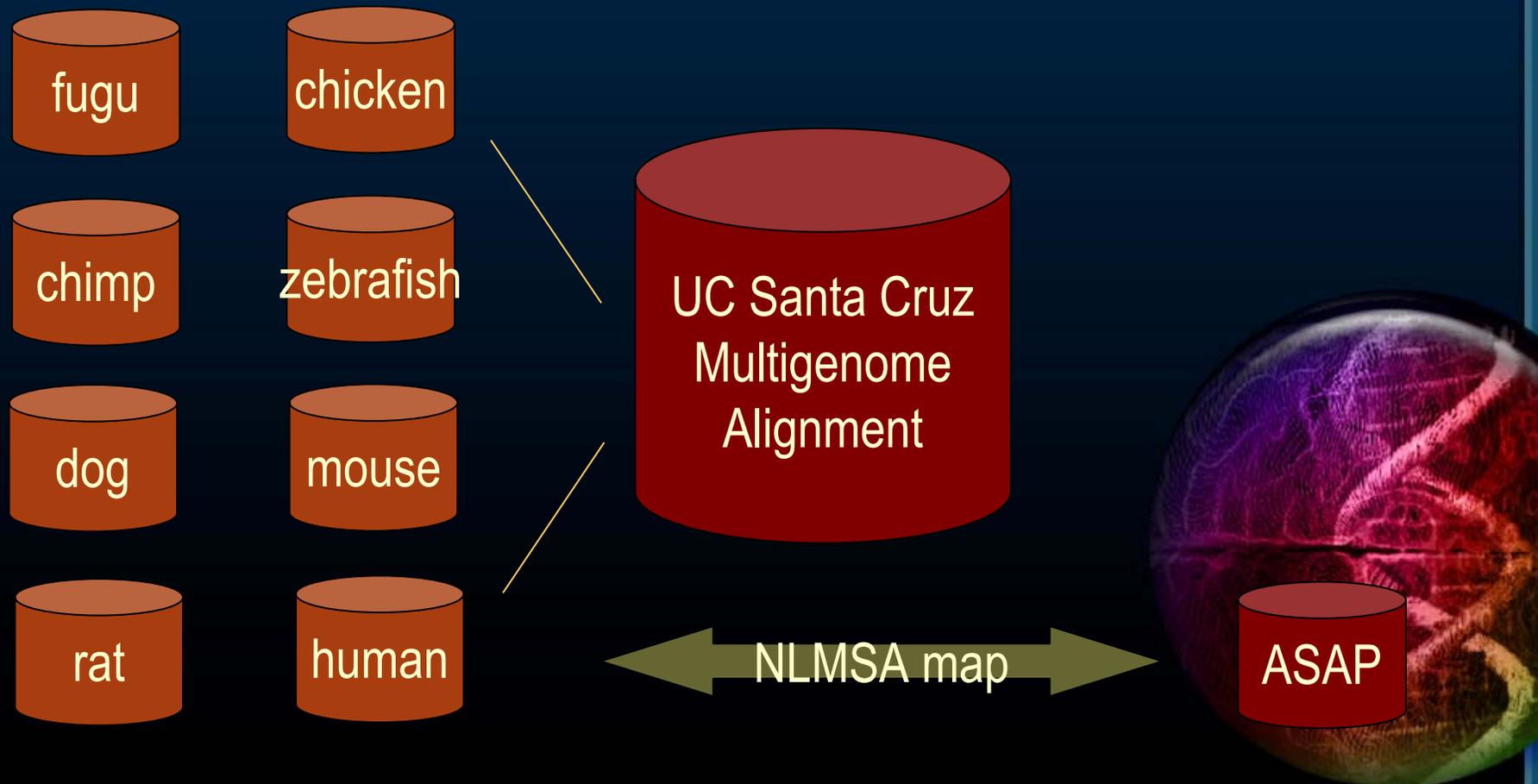
All of these can be represented as abstract intervals in 1D!

Alignment Query is the key operation in genome-wide sequence analysis



What does the "interval" I'm interested in align to?

Comparative Genomics Analysis of Alt. Splicing (17 genomes)



Genome Alignment Atlas Tools

- Pygr: query alignment as a graph
- NLMSA: highly scalable index for multigenome alignment & annotation
- Desktop PC can query terabyte scale multigenome alignments as easily as traditional single gene alignment.
- 0.2 msec/query (vs. 30 sec, MySQL)
- Included in Pygr software distribution.



Example: Genome-wide Exon Creation rates, measured in 17 animal genomes

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

High rates of exon
creation specifically
for alternatively
spliced exons.



New Tools for Atlases of Evolution

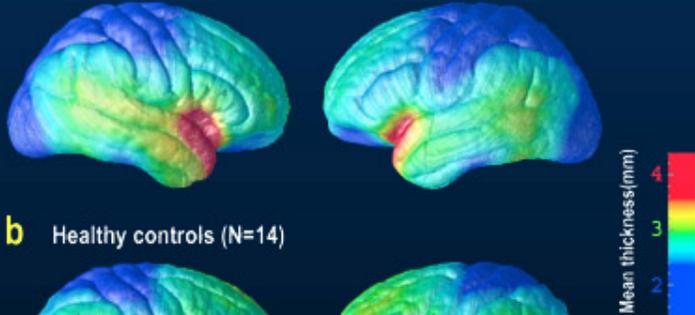
- BLASTgres: incorporate BLAST homology query directly in Postgres database (w/ Hsiao & Parker, UCLA)
- SNPindex: highly scalable graph database index for mutation covariation analysis (SNPs are nodes, edges are covariation relations).



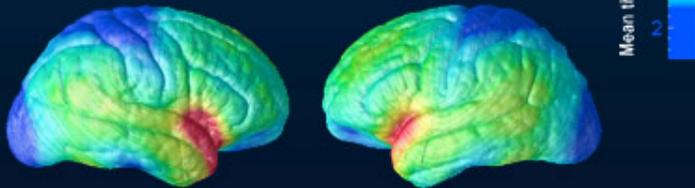
Atlases From Brains to Evolution: AIDS

Mean cortical thickness

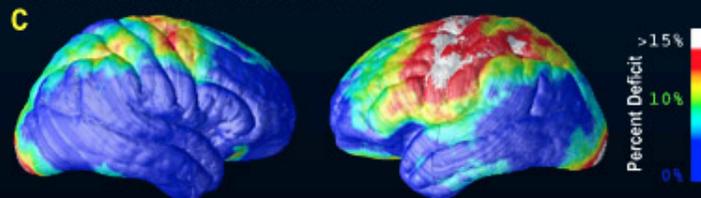
a HIV (N=27)



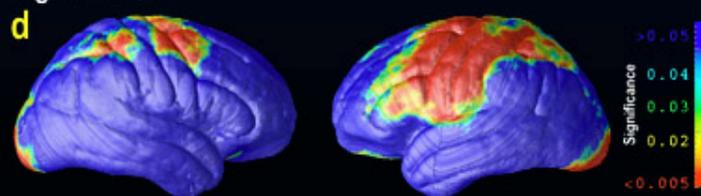
b Healthy controls (N=14)



Cortical thickness deficit in HIV

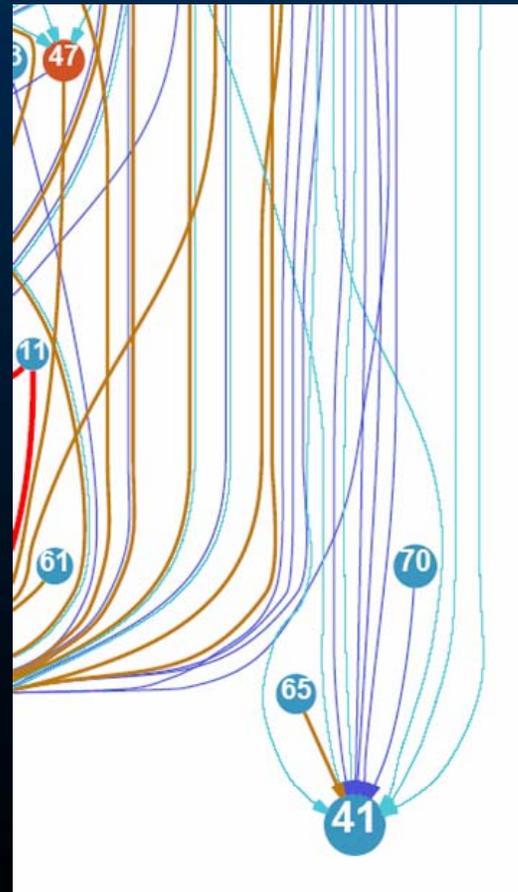


Significance



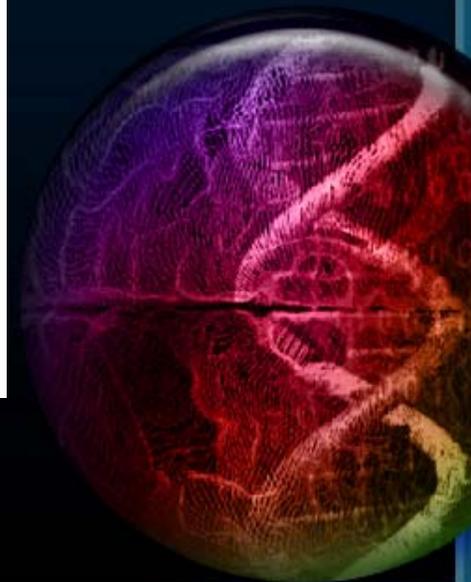
AIDS dementia
AIDS dementia

Can we map how HIV evolves?



HIV evolution

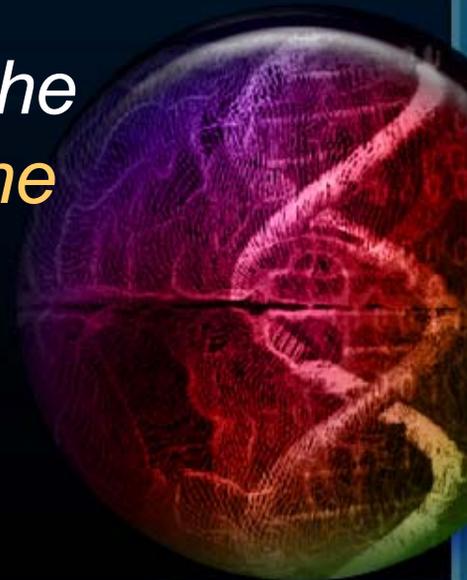
Model as a
snindex
graph
database.



Drug Resistance Evolution: We React to them, They React to Us... Stalemate

- E.g. a virus attacks us, so we develop a drug, so the virus evolves drug resistance mutations...
- Each step is a *reaction* to the enemy's *current state*, based on limited, *local information*, without considering *how the enemy will respond* to our actions *in the future*.

What if we had a *global atlas* of HIV's possible evolutionary *responses*?

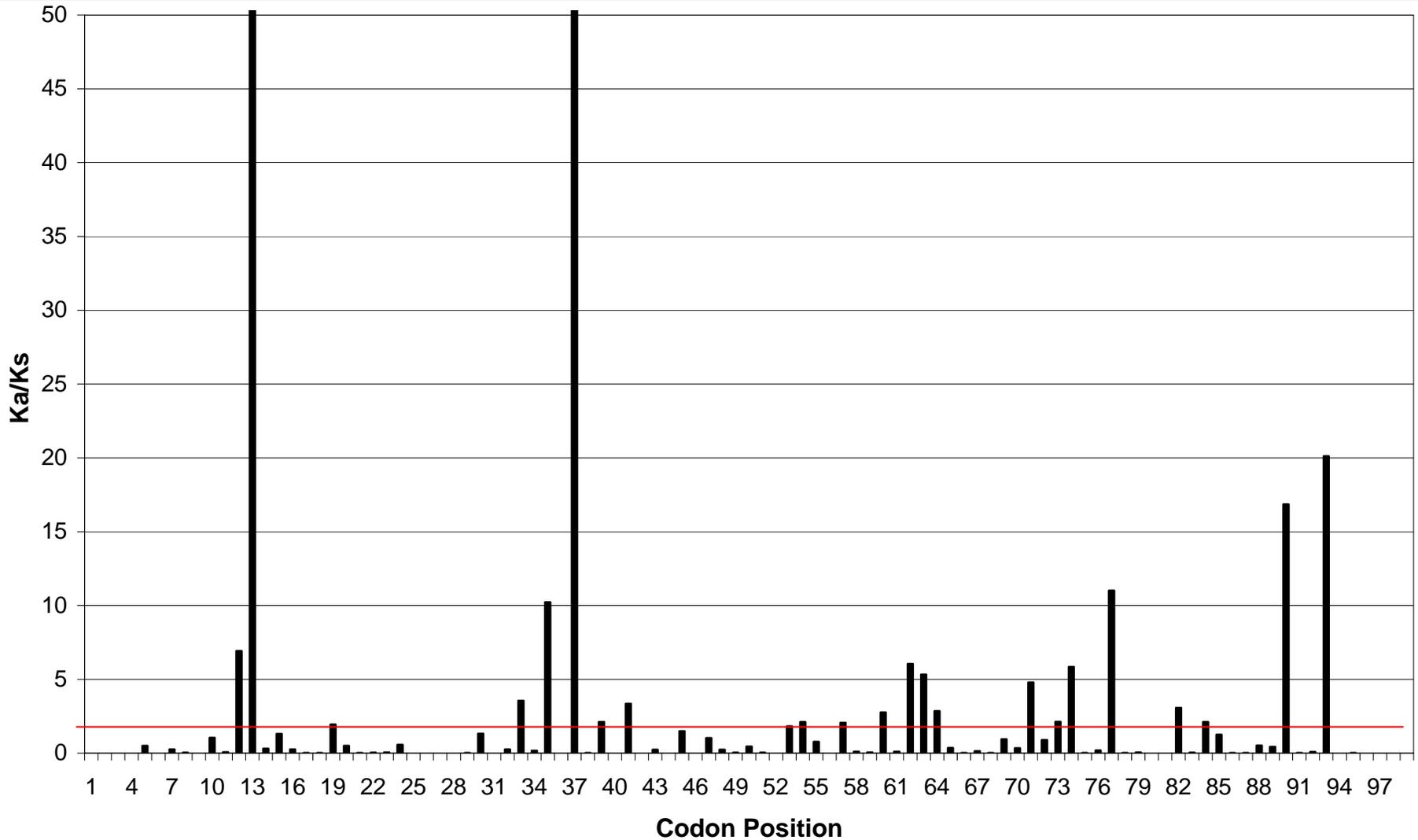


Selection Pressure Mapping: Build an Atlas of HIV Evolution

- *Selection pressure* measures whether an *amino acid mutation* is selected **for** ($Ka/Ks > 1$) or **against** ($Ka/Ks < 1$) by evolution, vs. *synonymous mutations*.
- Dataset: sequencing of 50,000 HIV clinical samples by Specialty Labs. Inc. 30-fold higher density of polymorphism information than human sequences.
- Goal: construct a **selection pressure map** of how HIV is evolving, where the virus is “going”, to evade our drugs.



HIV Protease Positive Selection



Positive selection mapping automatically discovers causes of drug resistance!

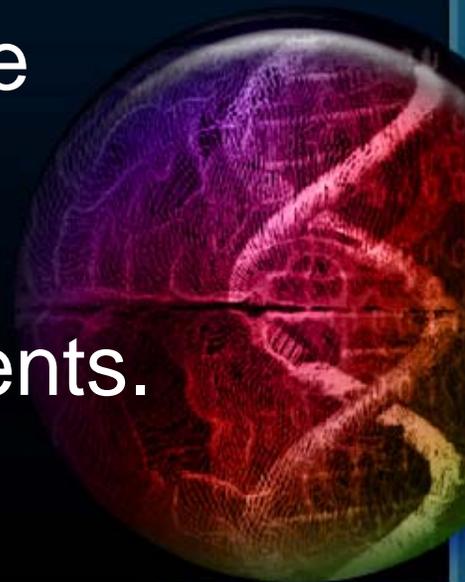
Positive Selection Mapping Identifies Drug Resistance Mutations

- Correctly identified 19 of 22 known drug resistance mutation positions.
- Compared with multi-year research process (clinical, biochemical, genetics) that was previously required, this analysis is completely automatic; works directly on output from sequencing machines.



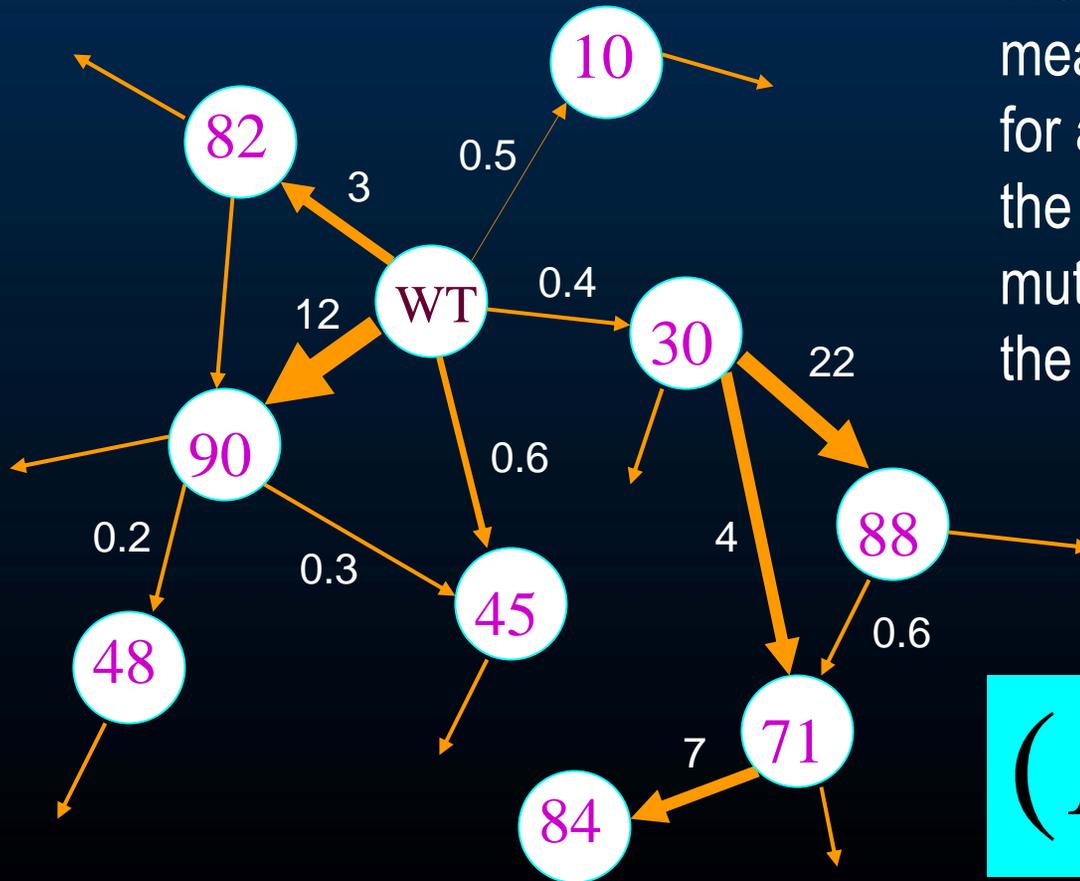
Build a Reaction Rate Diagram of HIV's Global Evolution

- A network diagram of the rates of transition between all possible genotypes. K_a/K_s is proportional to rate of increase of a mutation.
- Shows the speeds of all possible paths of evolution the viral population will follow, under the pressure of current drug treatments.



Conditional K_a/K_s Reveals Complete Mutation Network

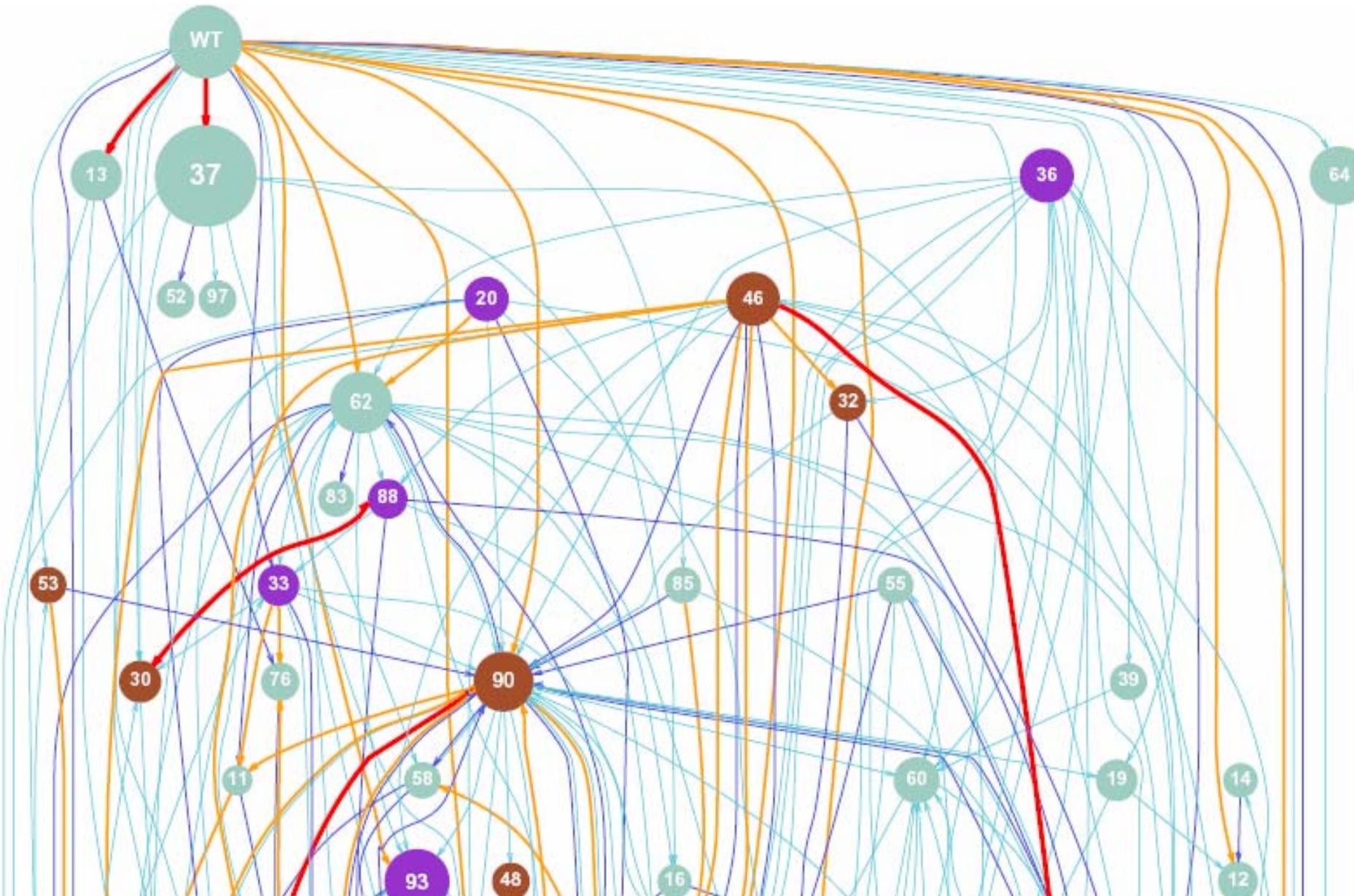
We can generalize K_a/K_s to measure the selection pressure for a mutation Y conditioned on the presence of a previous mutation X . We define this as the *conditional K_a/K_s* :



$$(K_a / K_s)_{Y | X}$$

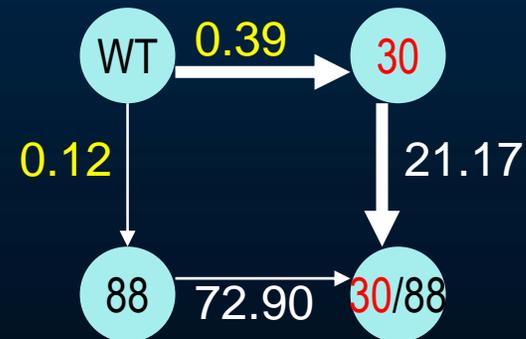
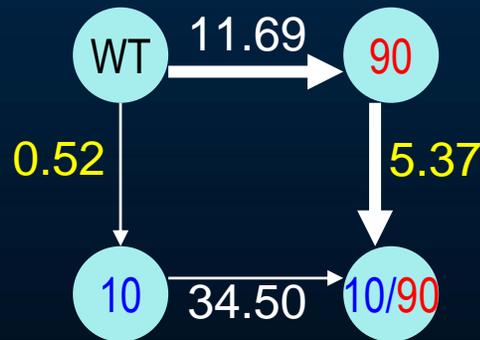
Each edge represents one conditional K_a/K_s value.

Fast Mutation Paths of HIV Protease



Different Paths to the Same Genotype can Differ in Speed

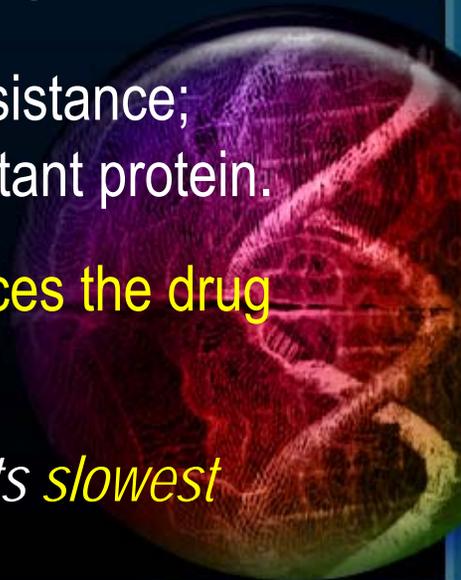
Chen & Lee, *Biol. Dir.* (2006)



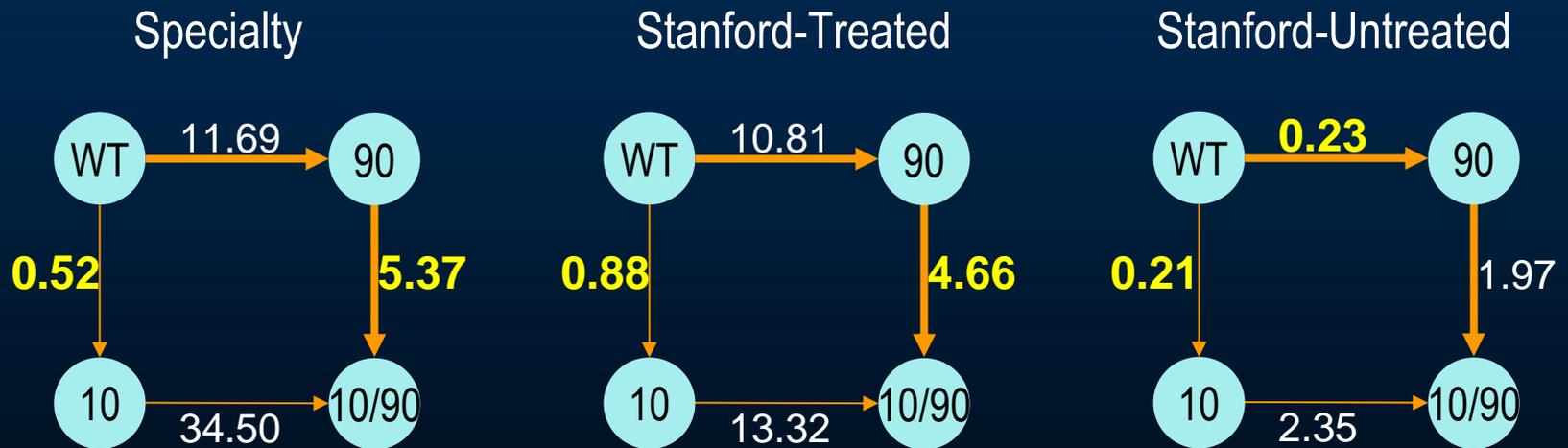
90 and **30** are mutations known to directly cause drug resistance; 10, and 88 are secondary mutations that stabilize the mutant protein.

Our analysis distinguishes them: faster path **first introduces the drug resistance mutation**, then the stabilizing mutation.

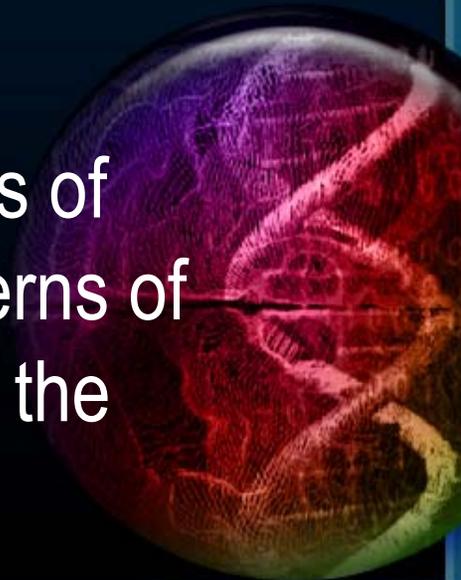
*NB: speed of a multistep path is generally controlled by its **slowest step**.*



Reproducible Results in Independent Datasets

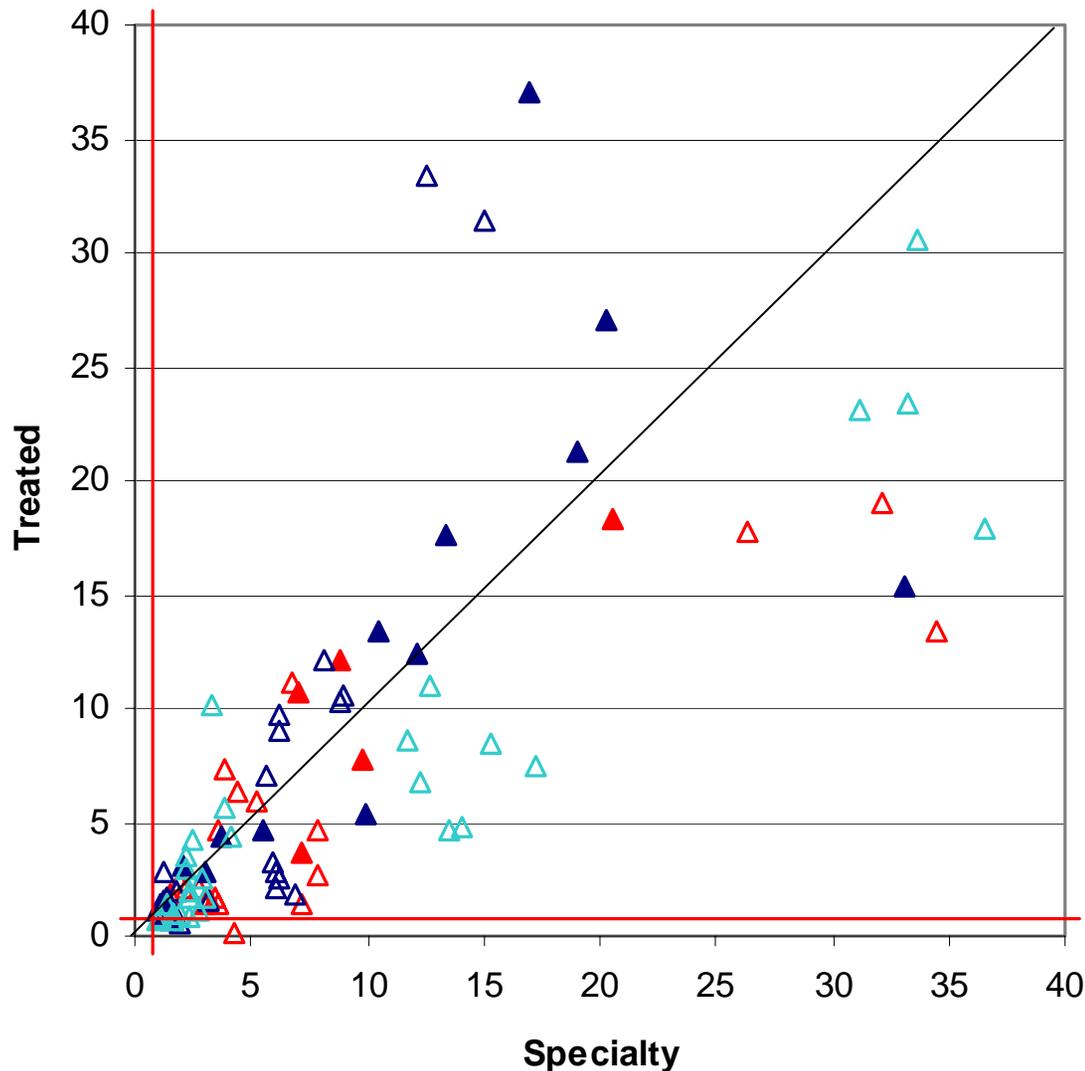


Highly reproducible in independent studies of different patients. They indicate real patterns of drug-associated selection pressure within the HIV population in the wild.



Quantitative Reproducibility

Conditional Ka/Ks: Specialty vs. Treated



- ▲ Primary drug resistance
- ▲ Accessory drug resistance
- △ Function unknown

For codons with sufficient counts ($N_{Xa} > 400$), the results match the Specialty results surprisingly well.

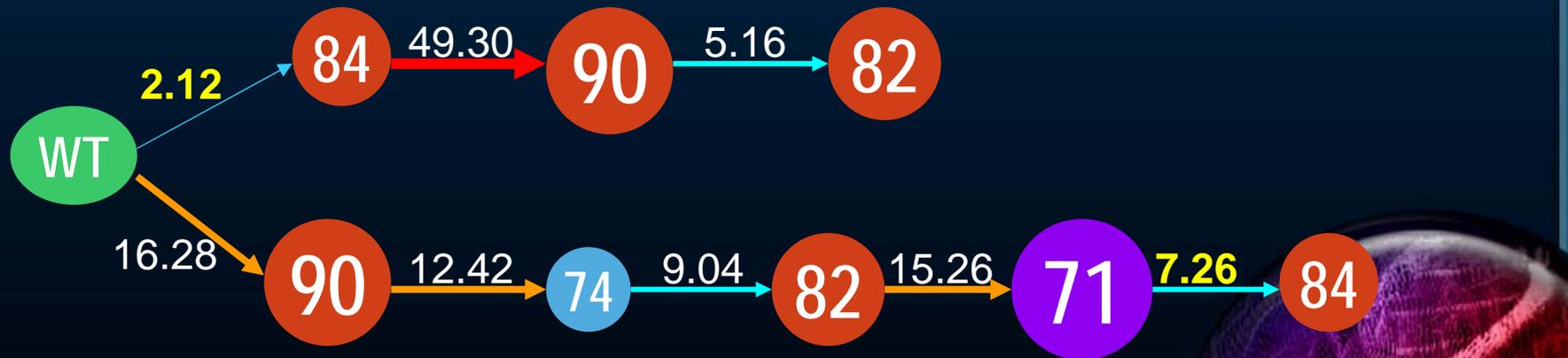
Danger: Fast Paths to Multi-Drug Resistance

- **Multiple resistance:** the combination of **three mutations** (at codons **82** (V82A/T/S), **84** (I84V), and **90** (L90M)) is resistant to most available protease inhibitors
- Rapid evolution of this triple mutant is a serious threat to individual treatment and to control of the global AIDS epidemic.
- Our map shows where the **fast paths** to this combination are. Don't want to go there!



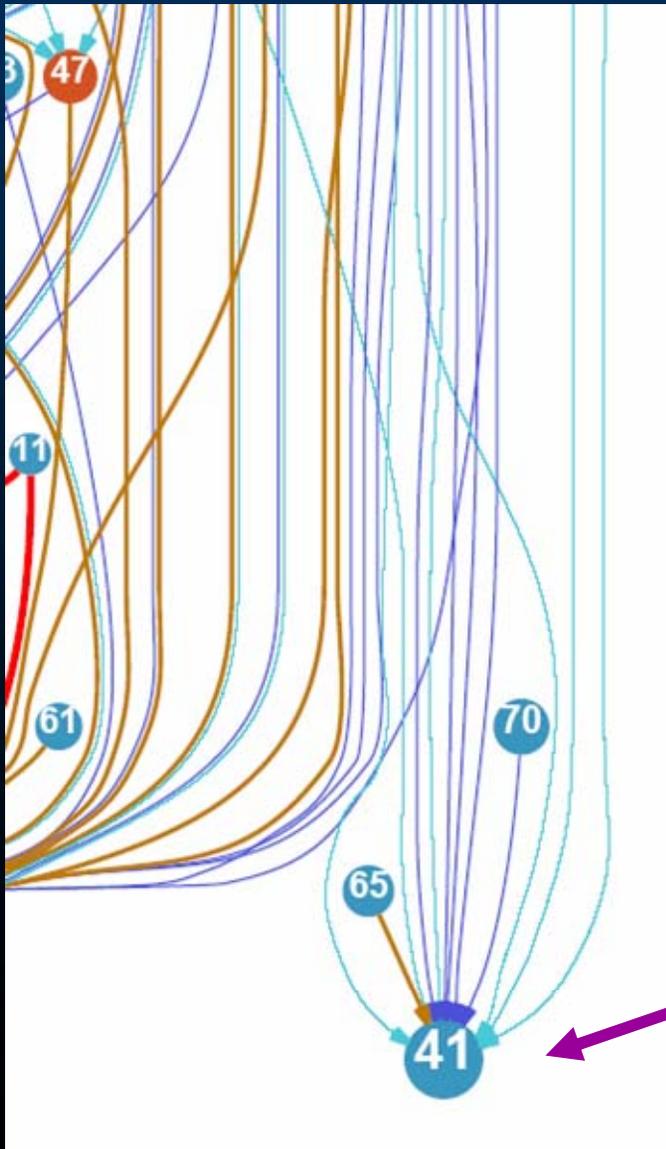
Reveals Accelerated Paths to Multi-Drug Resistance

The path that includes mutation at codons 74 and 71 is 3 times faster than the direct path, and 7 times faster in its first step:



Use the order in which drugs are given (which in turn can select for one mutation over another) to pick a slower path!

Kinetic Traps that Slow Evolution

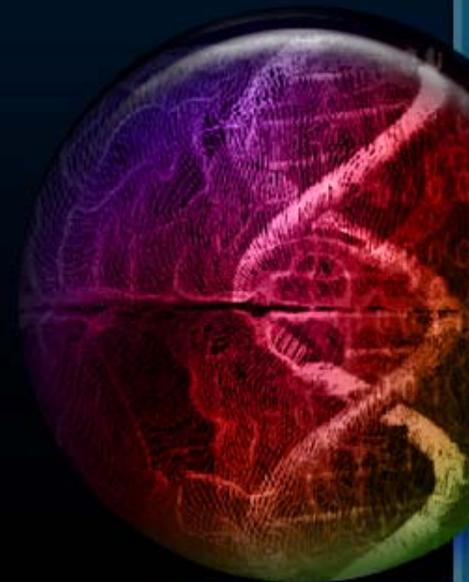


Kinetic Trap:
Many accelerated
paths to mutations
at 41, but no fast
paths to drug
resistant mutations
from there

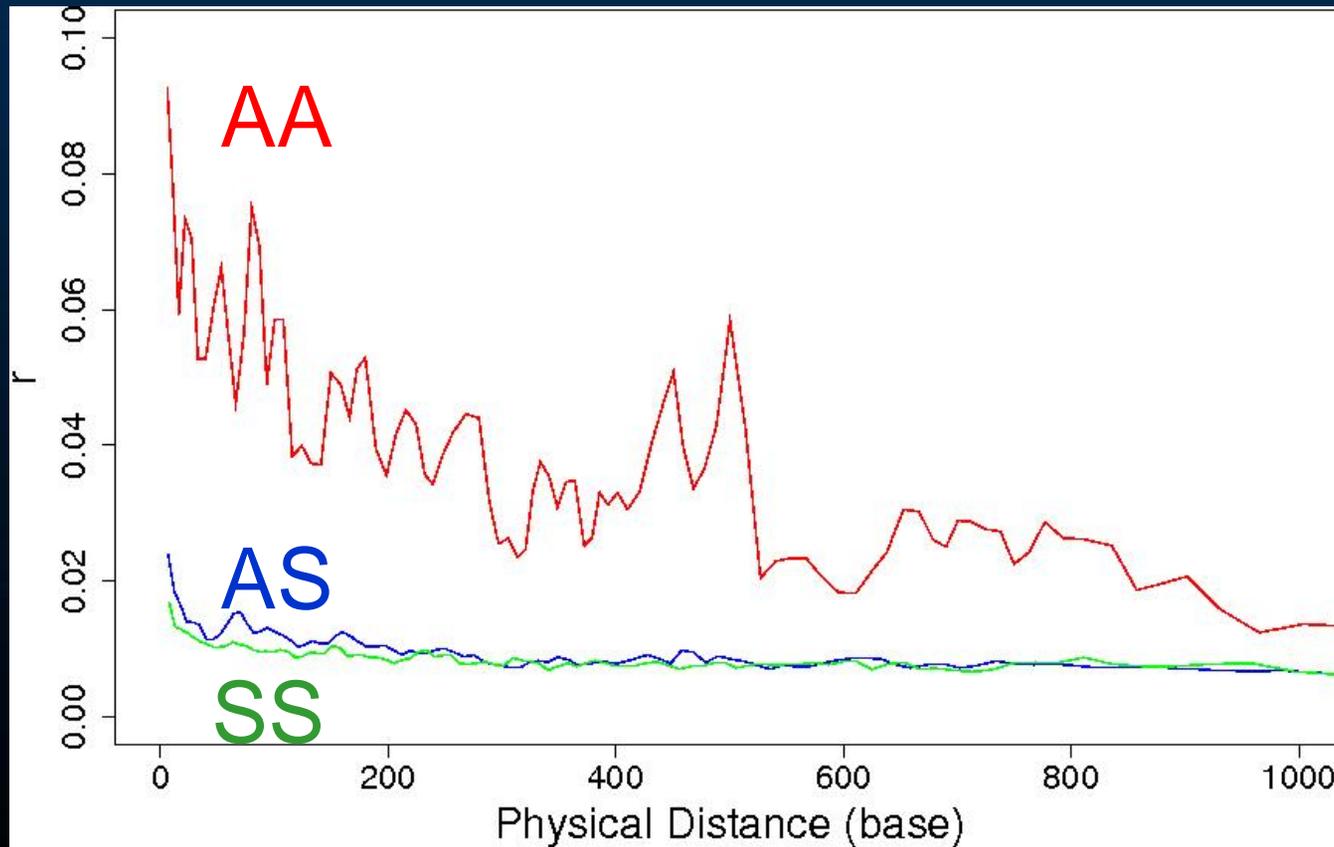


Tools for Separating Selection from Linkage using Synonymous Mutations

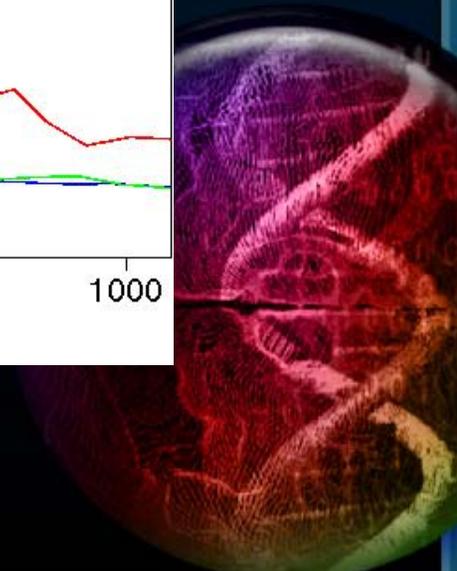
- Synonymous mutations should not be under selection, so they should give a direct readout of pure linkage.
- So, compare pairwise correlations of amino acid mutations (A) and synonymous mutations (S):
- AA, AS, SS



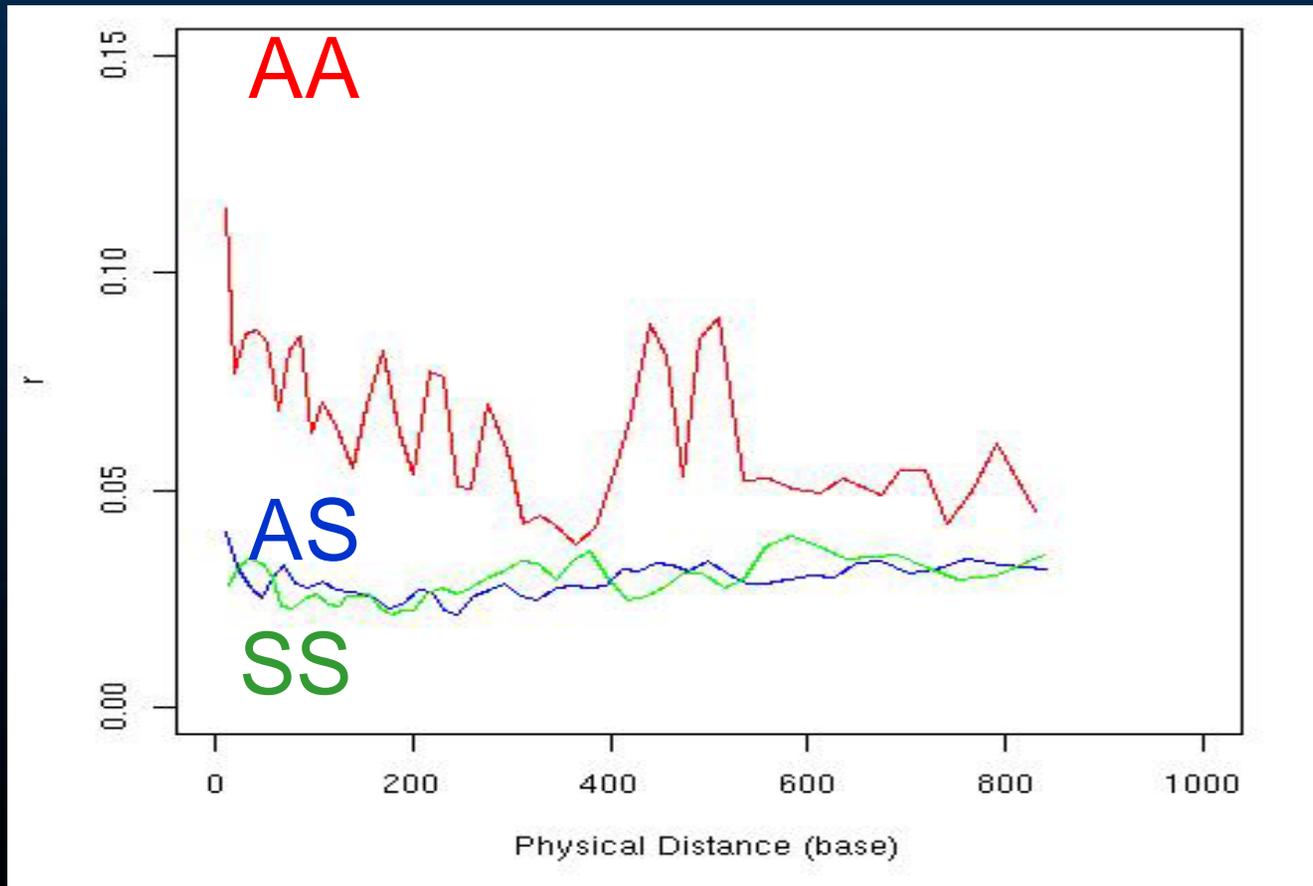
AA Covariation in HIV is due to Selection, not background LD (AS, SS)



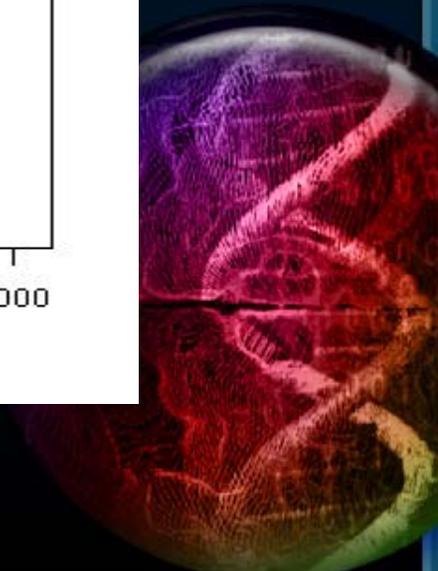
Specialty Dataset



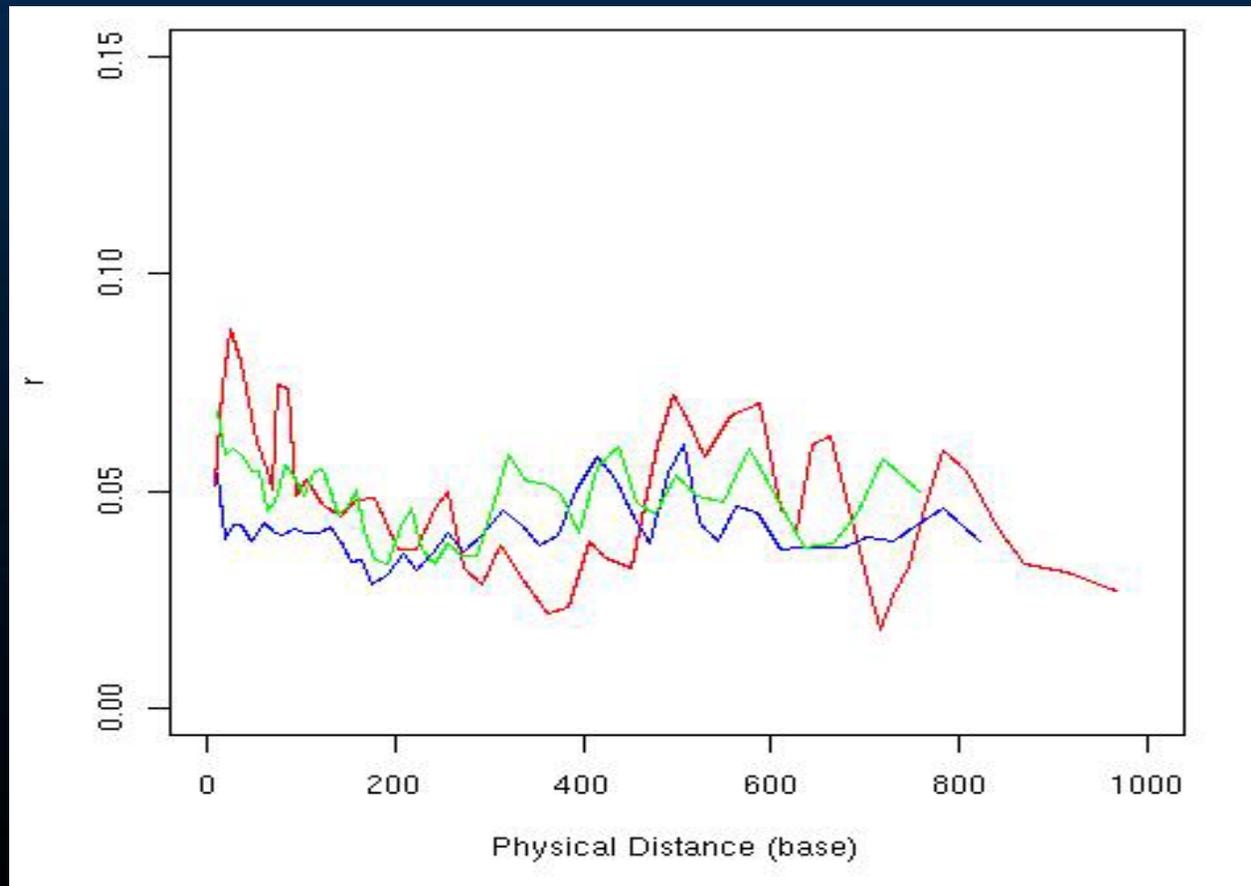
Reproducible Result in Independent Stanford-Treated Dataset



Stanford-Treated Dataset



Evidence of AA Selection Pressure Vanishes in absence of drug treatment



Stanford Untreated Dataset



A New Level of Strategic Intelligence

- A global picture of how HIV will respond in the future to our drug treatments.
- Ka/Ks velocities tell us where HIV population is *going*, detectable even while mutations still rare.
- Moreover, since these selection pressures are due to our actions (drugs), they are manipulable.
- Even slowing DR evolution two-fold could make a big difference for control of the epidemic.



HIV Positive Selection Database

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

<http://www.bioinformatics.ucla.edu/HIV/>

- Atlases of HIV drug resistance evolution from Specialty dataset.
- Analysis tools (snpindex).

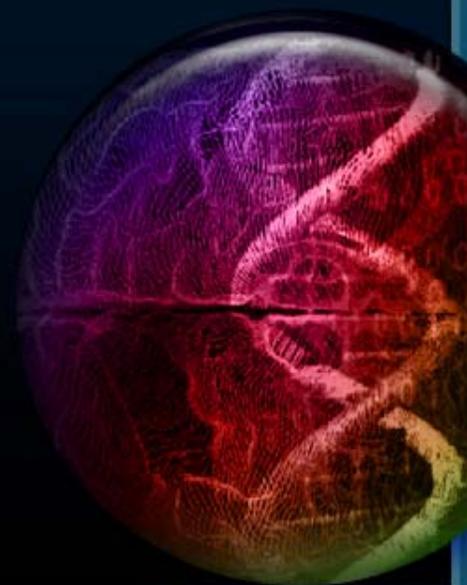


Acknowledgements

- **Pygr**: A. Alekseyenko, N. Kim, Z. Fierstadt
- **BLASTgres**: R.L. Hsiao, D.S. Parker
- **Snindex**: C. Pan
- **HIVdb**: L.M. Chen, Q. Wang, C. Pan, J. Kim
- **ASAP**: N. Kim

Generously supported by NIH, CCB

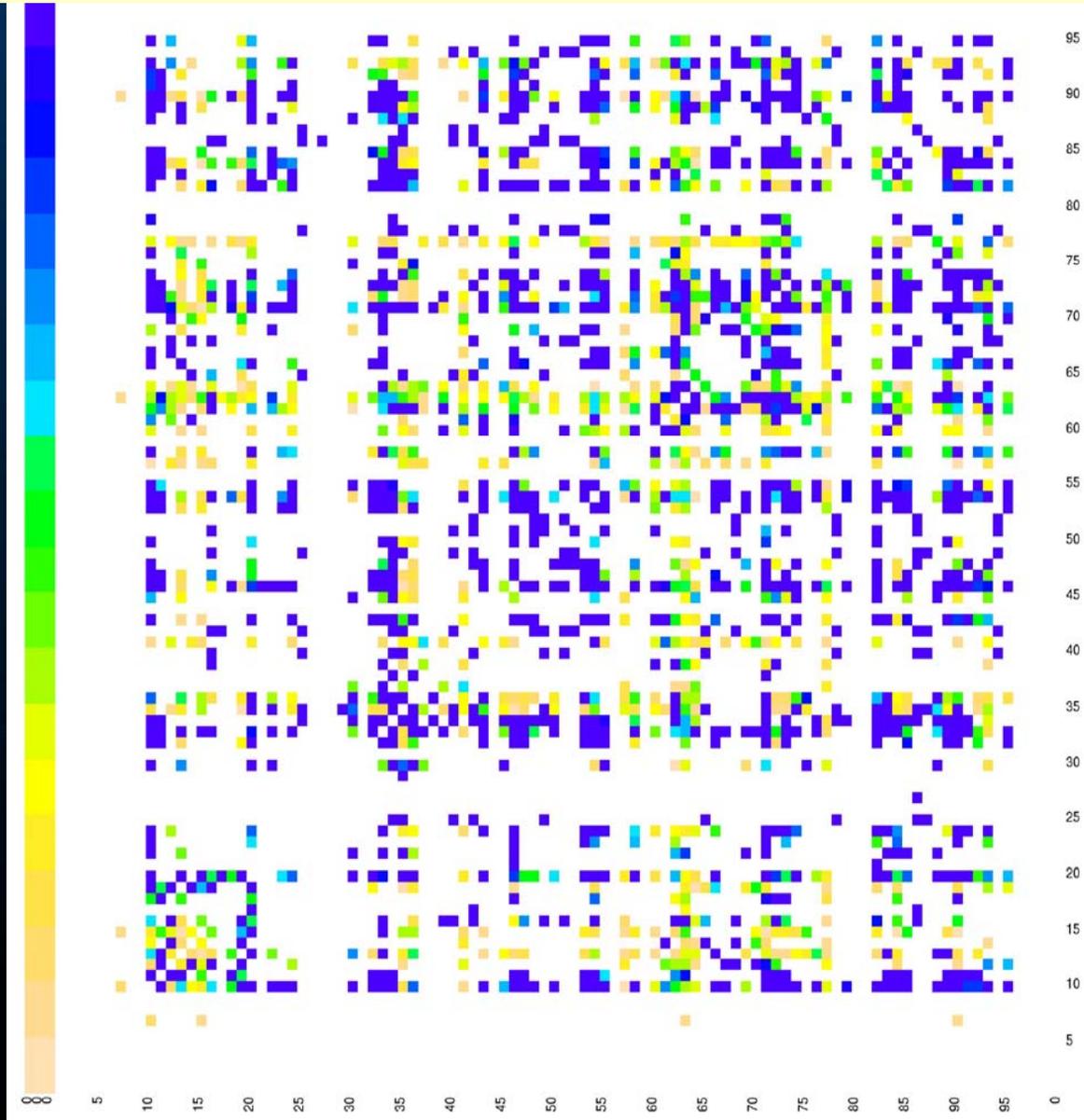




Mutation Associations (Protease, AA)

5

1

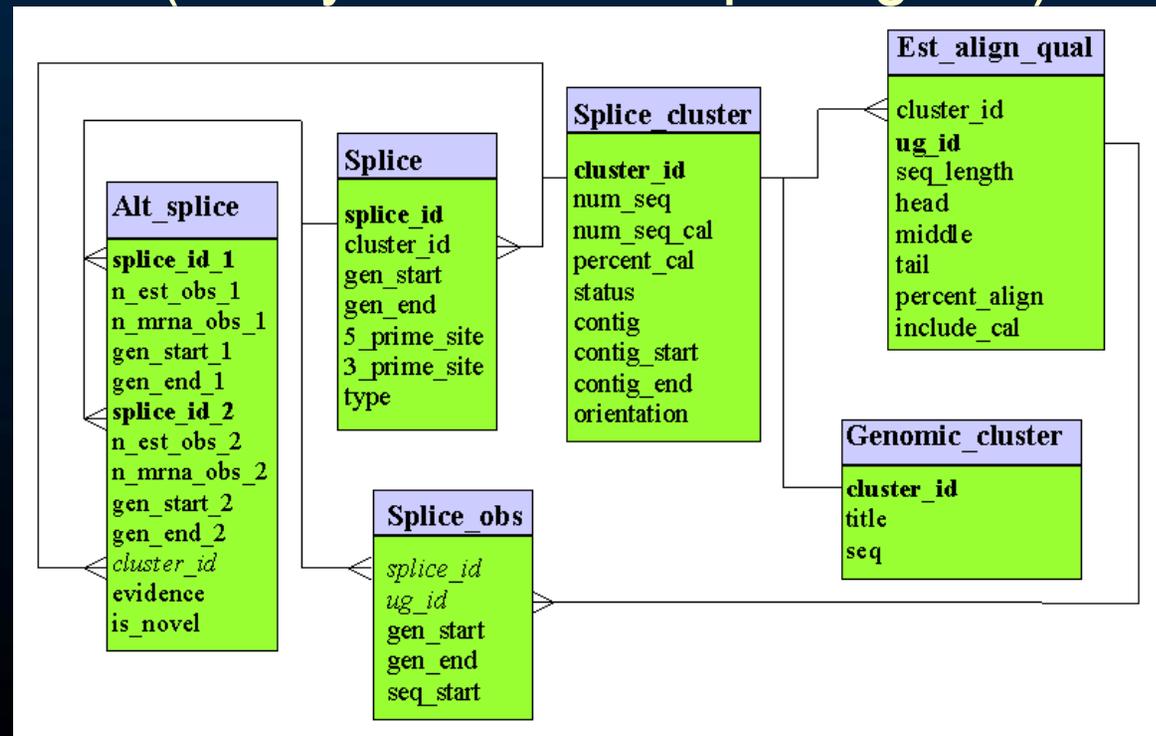


A



Hypergraphs are a General Model for Bioinformatics

Database Schema: (Entity-Relationship diagram)



Nodes: tables

Edges: entity-relationships (e.g. one-to-one, etc.)

Hypergraphs are a General Model for Bioinformatics

Dependency Graph: (e.g. make-rules)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Nodes: data types

Edges: make-rules

Selection Pressure is like an Evolutionary Velocity

For wildtype, synonymous mutant, and amino acid mutant allele frequencies $f_o=1$, $f_s=0$, $f_a=0$ initially, equal amino acid and synonymous mutation frequency λ , and reproduction rates r_o , r_a , after one unit of time

$$f_o \rightarrow r_o f_o (1 - 2\lambda); \quad f_s \rightarrow r_o \lambda f_o; \quad f_a \rightarrow r_a \lambda f_o$$

Assuming $\lambda \ll 1$, K_a/K_s and the normalized Δf_a will be:

$$\frac{K_a}{K_s} = \frac{f_a}{f_s} = \frac{r_a}{r_o}; \quad \Delta f_a = \frac{f_a}{f_o + f_s + f_a} \approx \lambda \frac{r_a}{r_o} = \lambda \frac{K_a}{K_s}$$

So initially the rate of change of the amino acid mutant allele frequency df_a/dt is proportional to the selection pressure K_a/K_s .

Multiple Independent Datasets

- **Specialty:** 50,634 samples representing a mix of treated and untreated patient samples from the U.S.
- **Treated:** 1,797 samples collected by Stanford University from patients with specific drug treatments
- **Untreated:** 2,628 samples collected by Stanford University specifically from untreated patients
- **Africa:** 399 African HIV-1 subtype C samples downloaded from Los Alamos HIV Sequence Database

