

Projects 2015

Erik Schultes

VOX: +31 064 244 8027

schultes@hedgehogresearch.info



Projects 2015

Sequenomics: Closing the gap between sequence & structure

- (1) 7-mer sequenome - data analysis
- (2) 7-mer sequenome - NK Model of rugged fitness landscapes
- (3) SeDEx platform - CS / Business ICT
- (4) SeDEx platform - Business ICT

Knowledge Dynamics: Knowledge representation & reasoning

- (5) Anatomy of a concept profile
- (6) Trend analysis

Sequenomics: Closing the gap between sequence & structure

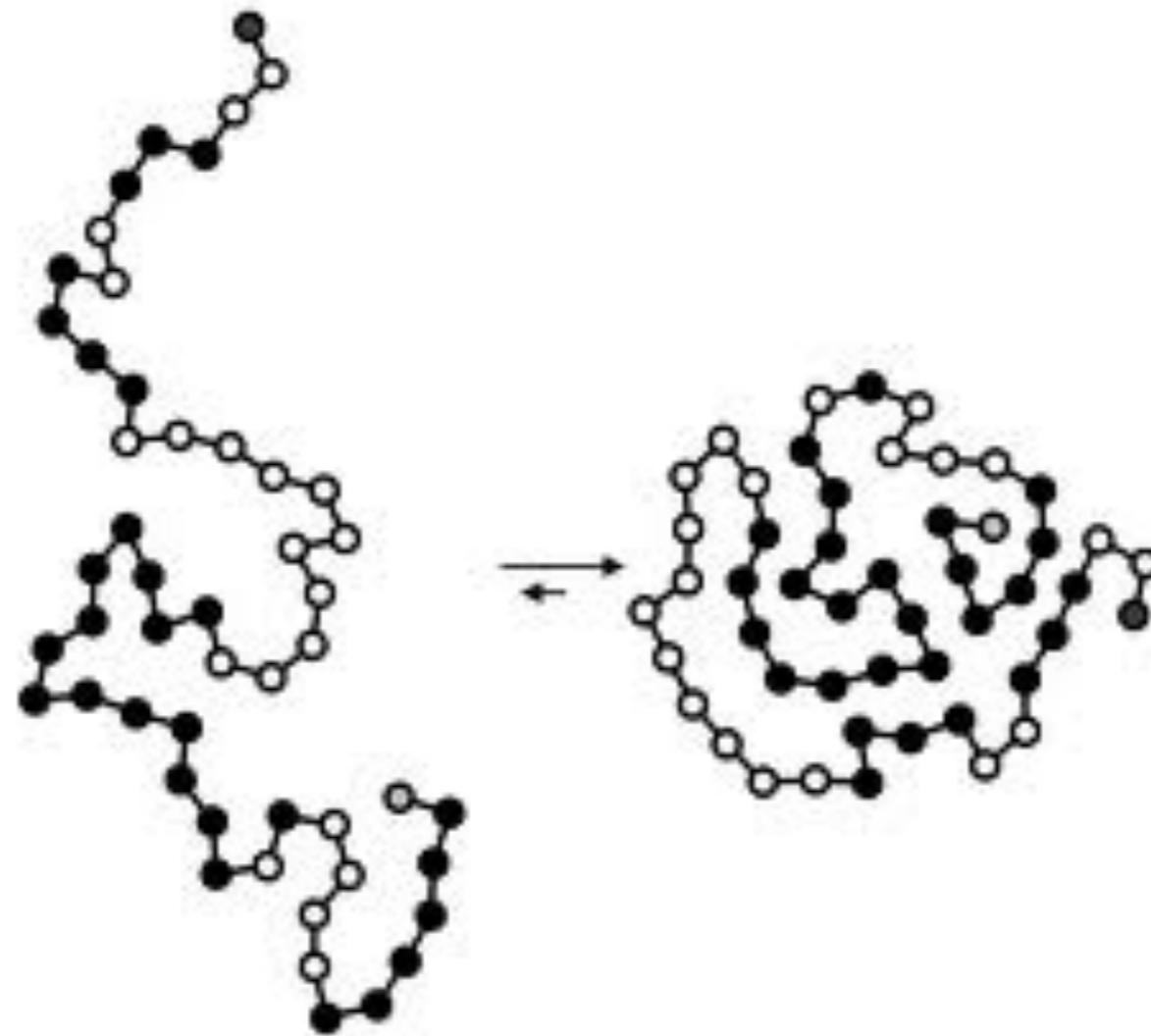
Sequence → Structure → Function

Sequenomics: Closing the gap between sequence & structure

Sequence → Structure → Function

Sequenomics: Closing the gap between sequence & structure

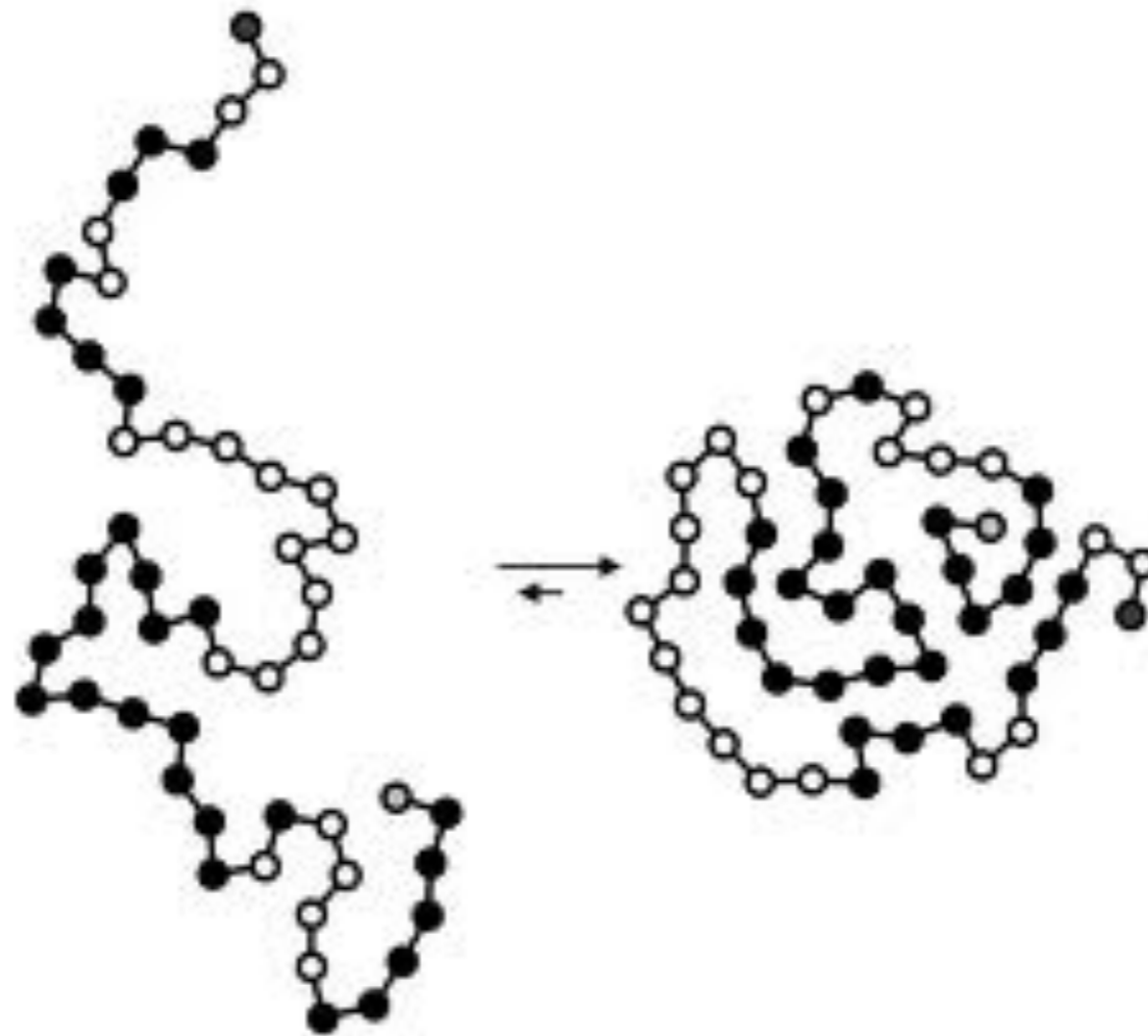
Sequence \rightarrow Structure \rightarrow Function



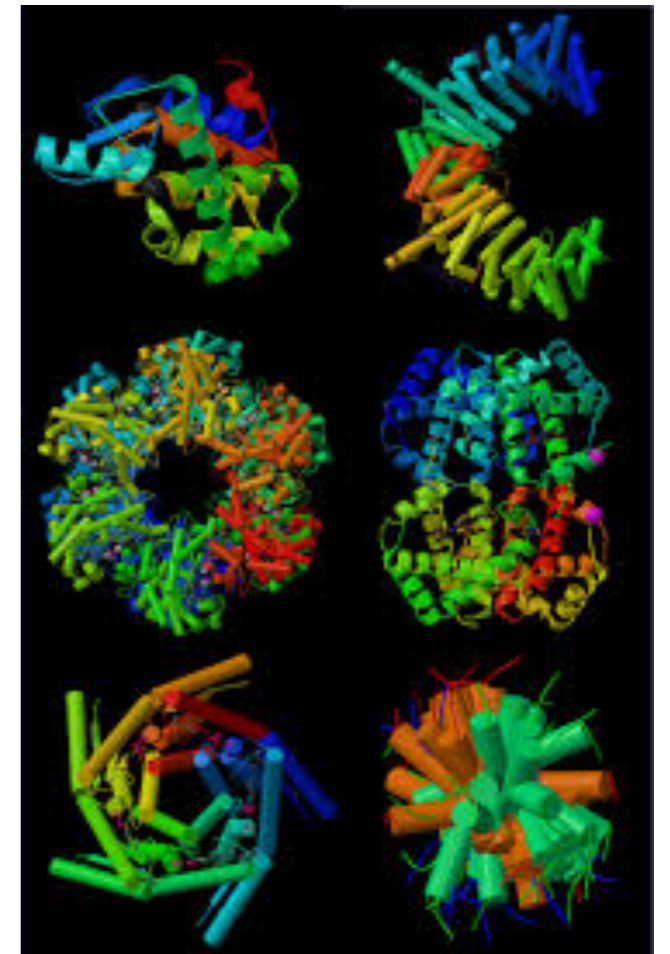
Sequenomics: Closing the gap between sequence & structure

Sequence → Structure → Function

easy



HARD

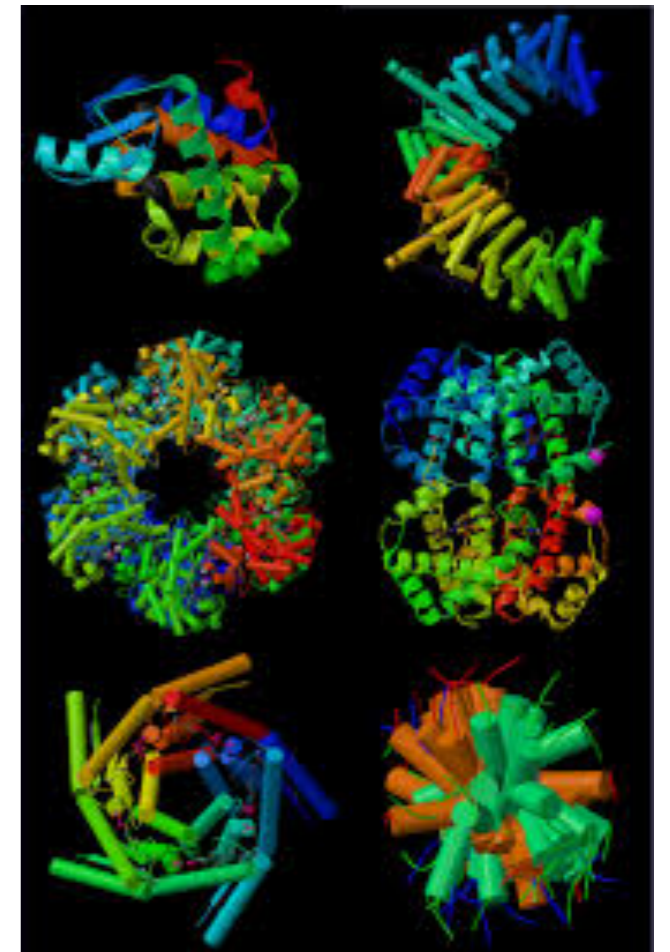
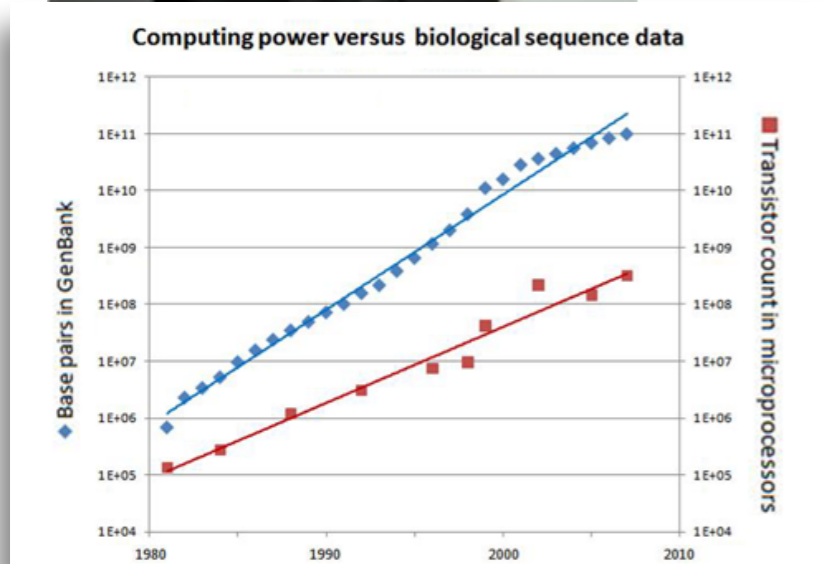
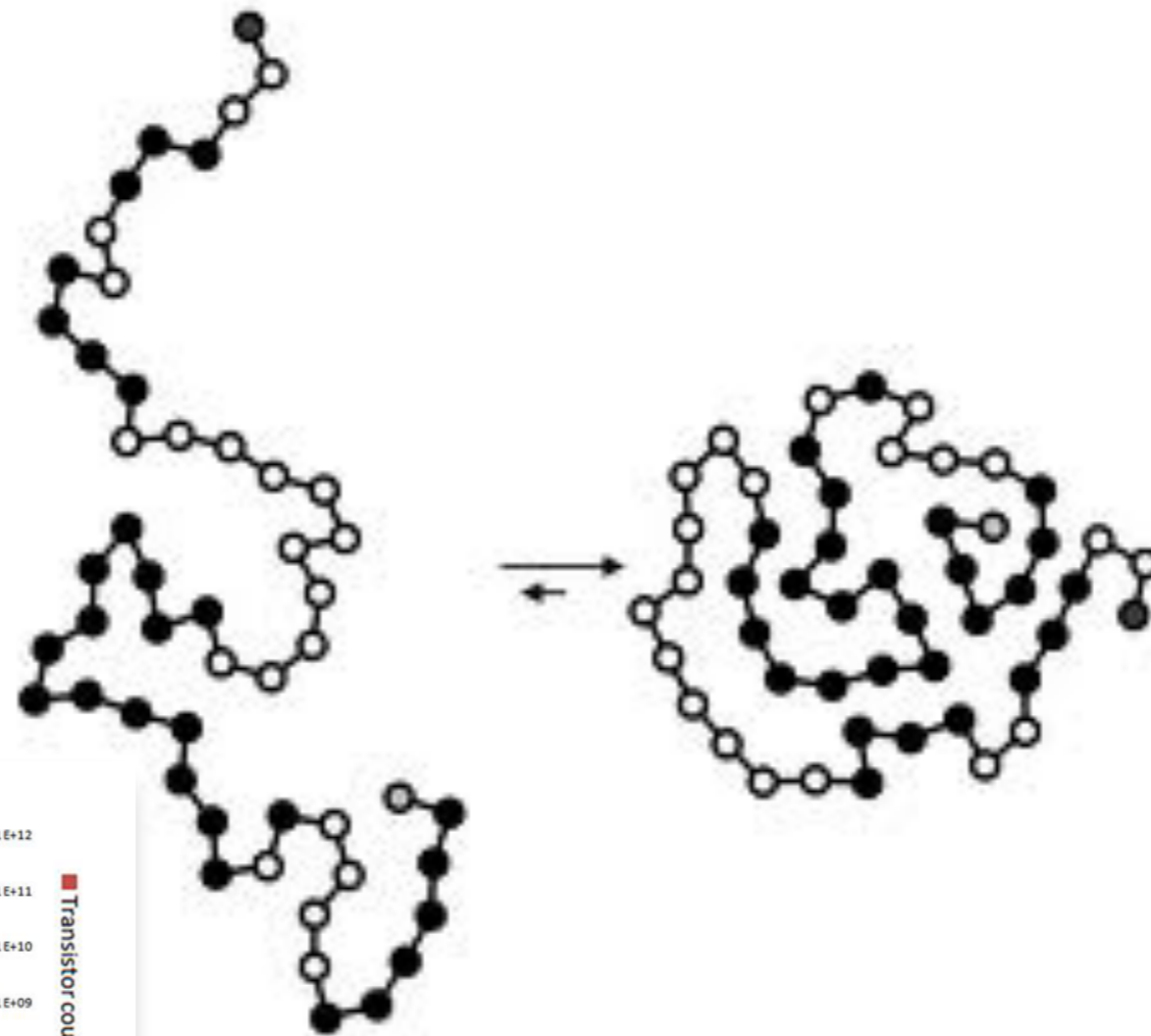


Sequenomics: Closing the gap between sequence & structure

Sequence → Structure → Function

easy

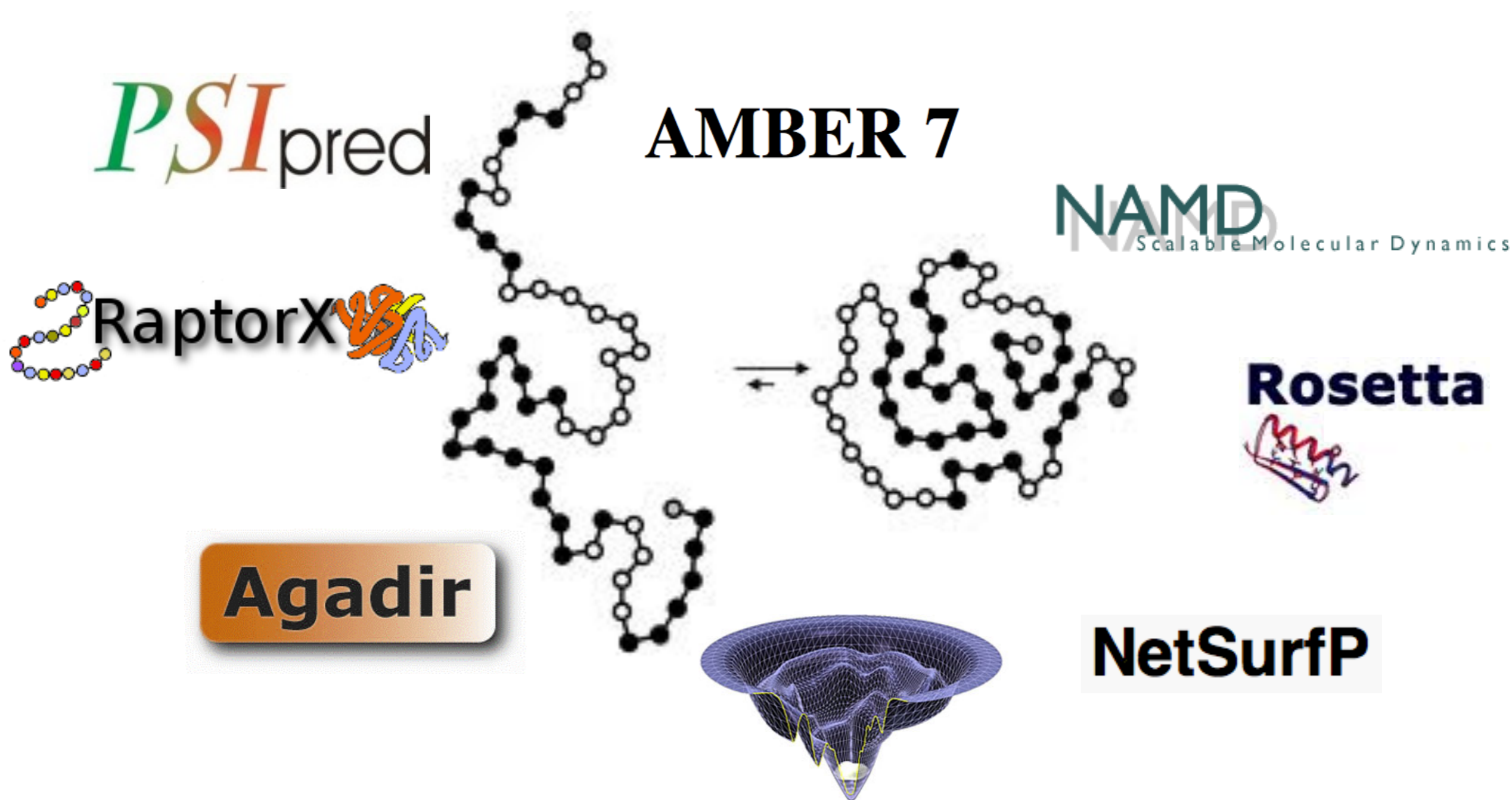
HARD



Sequenomics: Closing the gap between sequence & structure

Sequence → Structure → Function

In Silico Structure Prediction








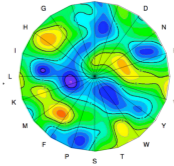
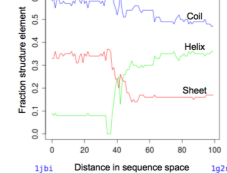






SeDEx

Sequenomics Data Exchange

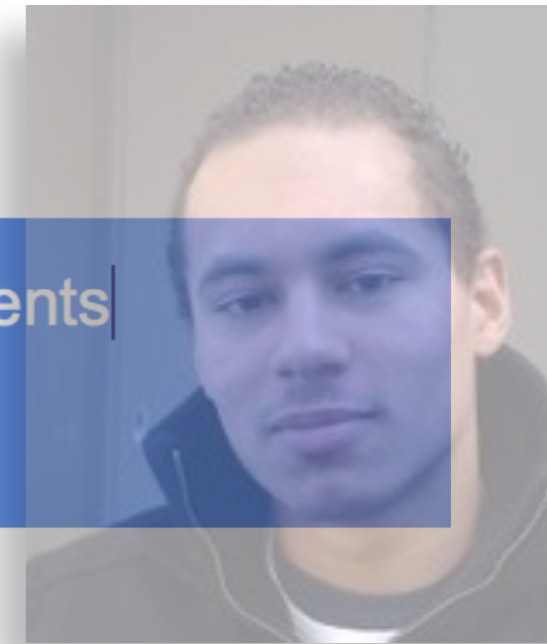
code \ data							
							
							
							
							
							
							

SeDEx

Sequenomics Data Exchange

code	data							
								
								
								
								
								
								

experiment



Welcome to the Sequenomics Data Exchange (SeDEX).

The SeDEX is an automated protein folding and structure prediction platform. SeDEX makes state-of-the-art structure prediction routine and easy.

Why Use SeDEX?

In the last 10 years it has become increasingly cheap and easy to obtain protein sequence information. But obtaining reliable structure information - through vital - remains tedious and expensive. SeDEX is designed to close the gap between amino acid sequence information & folded conformations.

How SeDEX Works

- SeDEX is a crowd sourced repository of protein sequence datasets & structure prediction algorithms.
- Registered users can post sequence data (whether a single sequence or many thousands) or install their structure prediction or protein folding software.
- The SeDEX automatically computes structural information for all sequences using all algorithms.
- SeDEX is an Open platform, although users can licence or restrict access to their data/code as desired.
- The SeDEX supports a simple API allowing its sequences, software and computed structures to be easily accessed as a web service.

Who developed SeDEX?

The SeDEX started as a idea pitched by Erik Schultes and was developed by Shamanou van Leeuwen as an internship project at the Leiden University Medical Center.

Logged in as:
shamanou van leeuwen
Logout

[Home](#) | [Algorithms](#) | [Datasets](#) | [Sequences](#) | [Experiments](#) | [Authors](#) | [Upload](#)

My Datasets

Huntington

My Algorithms

psipred
agadir
NAMD2

My Experiments

Huntington

Logged in as:
shamanou van leeuwen
Logout

[Home](#) | [Algorithms](#) | [Datasets](#) | [Sequences](#) | [Experiments](#) | [Authors](#) | [Upload](#)

Dataset - Huntington

Date created: 21/3/2014
Uploaded by: [shamanou van leeuwen](#)
License: 
Amount downloaded: 0
Funding source: lumc
Notes:

Sequences

This dataset contains 122 sequences.

Artificial sequences

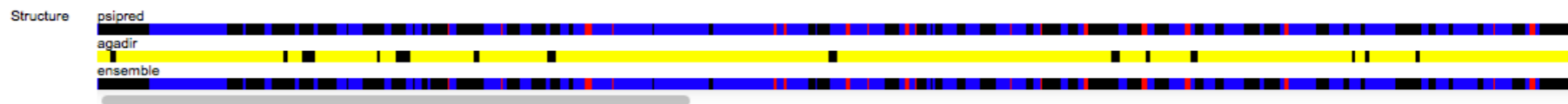
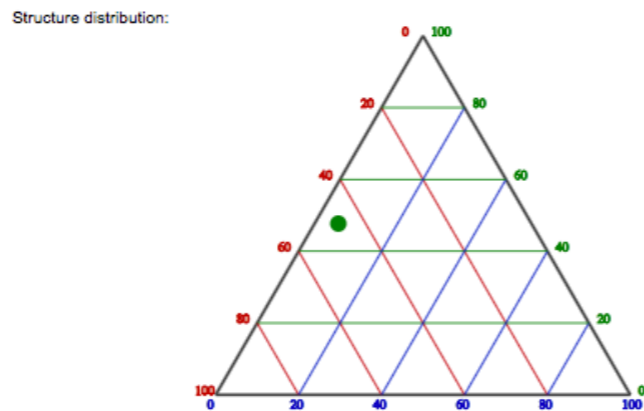
This dataset contains 121 artificial sequences.

[poly Q experiment 8 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 21 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 19 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 24 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 25 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 26 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 27 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 23 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 32 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 28 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 29 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 30 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 50 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 78 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 79 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 87 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 83 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 85 additional Qs HD_HUMAN Huntingtin](#)
[poly Q experiment 86 additional Qs HD_HUMAN Huntingtin](#)

Logged in as:
shamanou van leeuwen
Logout

Artificial Sequence

Name: poly Q experiment 7 additional Qs HD_HUMAN Huntingtin
Sequence length: 3149
Composition: D E F G A C L M N H I K T W V Q P S R Y
139 203 95 139 236 70 418 66 90 89 149 124 170 32 232 187 188 306 154 62



Dataset(s):
Huntington

Sequence: >poly Q experiment 7 additional Qs HD_HUMAN Huntingtin
QQQQQQMATLEKLMKAFESLKSFFQQQQQQQQQQQQQQQQQQPPPPPPPPPPPPQLPQPPPPQAPLQPPPPPPPPPPPPPPPPPPPPGPAVAEPLHRPKKELSA
TKKDRVNHCLTICENIYAQSVNRNPEFQKLLGIMELFLCSDDAESDVVMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAPLA
HLVRFQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSIQHSRRTYFYYSWLLNV
LLGLLVPVEDEHSTLLILGVLLTLRYLVPLLLQQVKDTSLKGSPGVTREKEMEVSPSAEQLVQVYELTLHHTQHQNHNVTGALELLQQLFRTPPELLQ
LTAVGGIGQLTAAKESGGRRSRSGSIVELIAGGGSSCSFVLSRKQKGVLLGEEEALEDSDSSRSVDSSSALTASVKDEISGELAASSGVSTPGSAGHDI
ITEQPRSQHTLQADSVLDASCDLTSSTADGDDEEDILSHSSSQVSAVPSPAMDNDGTQASSPIDSSQTTEBGPDSAVTPSDSSEIVLDGTDNQYLGLQ
IGQPQDEEEATGILPDEASEAFRNSMALQQAHLKXNMSHCRQPSDSSVDFVLRDEATEPGDQENKPCRIGDQGSDDDSAVLVHCVRLLSASFLL
TGGKNVLPDRDVRVSVKALALSCVGAVALHPEEFFSKLYKVPLDTEYEPBQVSDILNYIDHGDPQVRGATAILCGTLCISILSRSRFHVGMWGTI
RTLGTNFTSLADCIPLLRKTLKDESSVTCCLACTAVRNCVMSLCSSESYSELGLQIIIDVLTLRNNSYWLVRTELETLAEIDFRLVDFLEAKAENLHRGA
HHYTGLLKLQERVLNNVVIHLLGDEDPRVRHVAASLIRLVPKLFYKQDQADPVAVARDQSSVYLKLLMHETQPPSHFSVSTIRIYRGYNLLPSIT
DVTMENLSRVIAAVSHELITSRALTFGCCEALCLLSTAFFVCVWSLGHGCVPPPLSASDESRRKSCVGMATMILTLSSAWFPLDLSAHQDALILAG
NLLAASAPKSLRSMWASEEENPAATKQEEVWPALGDRALVPMVEQLFSLHLLKVINICAHVLDVAPGPAIKAALPSTLNPPSLSPIRRKKEKEPEGEQA
SVPLSPKKGSEASAASRQSDTSGPVTTSKSSSLGSFYHLPYKLHVDLKAHANYKVTLDLQNSTEKFPGGFLRSALDVLQSQILELATLQDQIGKCVVEIL
GYLKSCFSREPMMATCVQQLKTLFQGNLASQFDGLSSNPSSKQGRARQLGSSSVRPGLYHYCFMAYTHFTQALADASLNMVQAEQENDTSGWFDVL
QKYSTQLKNTLSTVTKNRADKNAIHNHRLFEPLVIKALKQYTTTCVQLQKQVLDLLAQVLQVLRVNYCLLSDQVFIGFVLKQFYEIVGQFRESBAII
PNIFFLVLLSERYHKSQIIGIPKIQIQLCDGIMASGRKAVTHAIPALQPIVHDLFVLRGNTKADAGKELETQKEVVVSMMLRLIQHVQVLEMFILVLLQ
CHKENEDKWKRLSRQIADIILPMLAKQMMHDSHEALGVLNLFPEIILAPSSLRPVDMLLRSFVTPNTMASVSTVQLWISGILAILRVLISQSTEDIVLS
RIQELSFSPYLSICTVINRLRDGDSTSTLEHSESKQIKNLPETFSRFLQLVGIILLEDIVTKQKVMSEQQHTFYCQLGTLMLLIHIFKSGMFRV
ITAAATRLFRSDGCGSFFYTLDSNLNRARSMITTHPALVLLWCQIIILLNHTDYRWAEEVQVTPKRHSLSTKLLSPQMSGEEEDSCLAAGLGMCNREIV
RRGALILFCDYVCQNLHDSHETWLVNHIQDLISLSHEPPVQDFISAVHRNSAASGLFIAIQSRNCENLSTPTMLKKTQCLEGHLSQSGAVLTLYVD
RLCTPPFRVLRMVDILACRRVEMLLAANLQSSMAQLPMEELNRIQEYLSQSGLAQRHQRYSLLDRFLRTMQDLSLSPFPVSSHPDGDGHSVLETVS
PDKDWYVHLVKSQCWTRSDSALLEGAELVNRIPAEOMNFMNNEPNLAPCLSLGMSEISGGKQKALFEAREVTLARVSGTVQQLPAPHVHFQPEL
PAEPAAVWSKLNDFGDAALYQSLPTLARAALQYLVVSKLPSHLHLPPEKEKDIVKVVVATLEALS WHLHEQIPLSLDLQAGLDCCLALQPLGLWSV
VSTEFVTHACSLIYCVHFLEAVAVQPEGQLSPERRTNTPKAISEEEVEEDPNTQNKYIITACEMVAEMVESLQSVLALGHKRNSSVGFALTPLLRN
IIISLARLPLVNSYTRVPLVWKLGS PKPGDFGTAFPEIPEVFLQEKVEKFEIYRINTLWTSRTQFEETWATLLGVLVTQPLVMEQEBSPPEEDTE
RTQINVLAVQAITSLVLSAMTVFVAGNPAVSCLEQQPRNKPKALDRFRGRKLSIIRGIVEQEIQAMVSKRENIAETHLYQAWDPVPSLSPATTGALISH
EKLMLQINPERELGSMYSYKLGQVSIHNSVLGNSITPLREEEWEDEEEDADAPAPSPPTSPVNRKRRAGVDIHSQSQFLELYSRWILPSSARRTPA
LLISEVVRSLLVSDLFTERNQFELMYVTLELRRVHPSSEDEILAQYLVPAACKAAAVLGMKVAEPVSRLESTLRSSHLPSRVGALHGVLVYVLEDL
LDDTAKQLIPVDSYLLSNLKGIAHCVNIHQQHVLVMCATAFYLIENYPLDVGPEFSASIIQMCQVMSGSEESTPSIYHCALRGLERLLLSEQLSRL
DAESLVLKSVDRVNVHSPHRAALGLMLTCMYTGEKVS PGRSDFNPAAPSDSEIVAMERVSVLFDRIKGFPCERAVVARILPQFLDDFFPQDIM
NKVIGBFLSNQQPFYQFMAVYVYKVFQTLHSTGSSMVRDWMLSLNSPTQRAPVAMATWSLSCFFVSASTSPWVAAILPHVISRMGMKLEQVDVNLFCLV
ATDFYRHQIEEELDRRAFQSVLEVVAAPGSPYHRLLTCLRNVHKTTC

Purpose of the SeDEx

Biologist

- Make structure prediction easy - just upload a data file
- Repository of sequences (natural, artificial)
- Repository of algorithms (automatic, versioning)
- Repository of structure predictions (experiments)
- Accuracy: Consensus predictions

Computer Scientist

- Market place for software
- Head-to-head comparison of software performance on different sequences
- Where are software consistent? where are they not?

Access to the SeDEx

Browsing

- View datasets and sequences
- View software profiles
- No download capability

Registered User

- Unique identifier
- Funding
- Skype ID
- Upload / Download datasets
- Upload / Download code (via request to Admin)
- Search authors



Use of the SeDEx

Computational Load Issues

- Secondary structure predictions are fast (and likely free)
- 3D structure predictions can be very slow (and will cost the user)



Licensing Issues

- Author controlled
- Open / Closed
- Free / Fee



Nanopublication

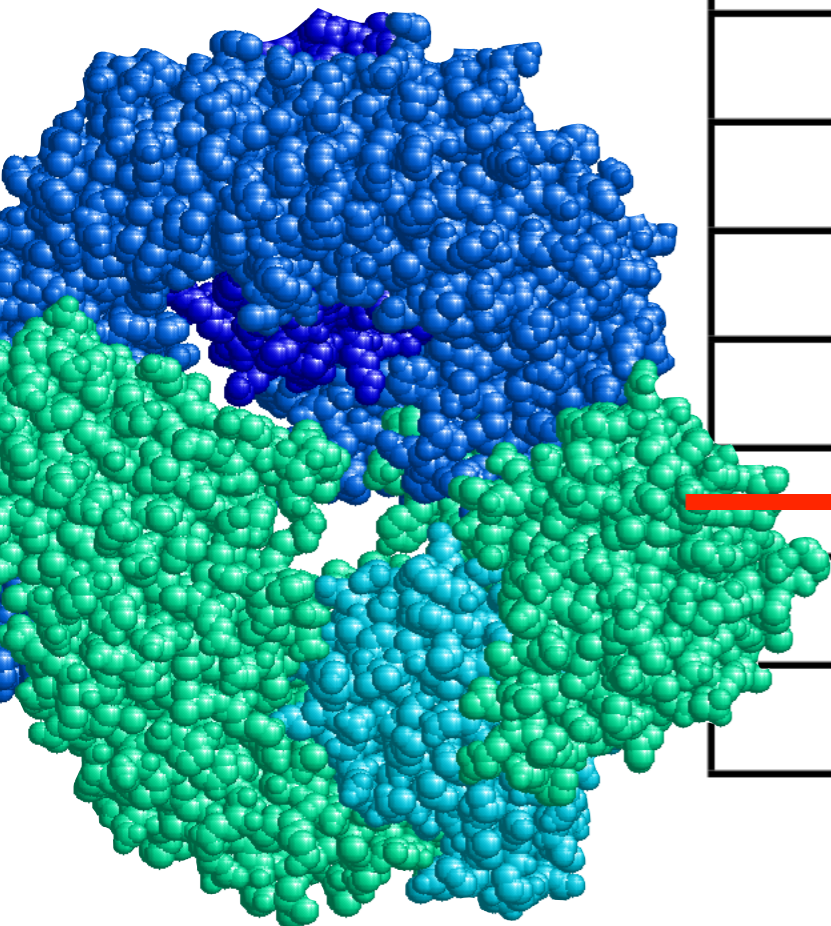
- Datasets are citable
- Software is citable
- Individual sequence-structure combinations are citable

The Sequenome

sequence length	possible sequences
2	4.0E+02
3	8.0E+03
4	1.6E+05
5	3.2E+06
6	6.4E+07
7	1.3E+09
8	2.6E+10
12	4.1E+15
40	1.1E+52
80	1.2E+104
160	1.5E+208
300	2.0E+390
320	2.1E+416
1000	1.1E+1301

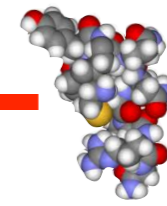
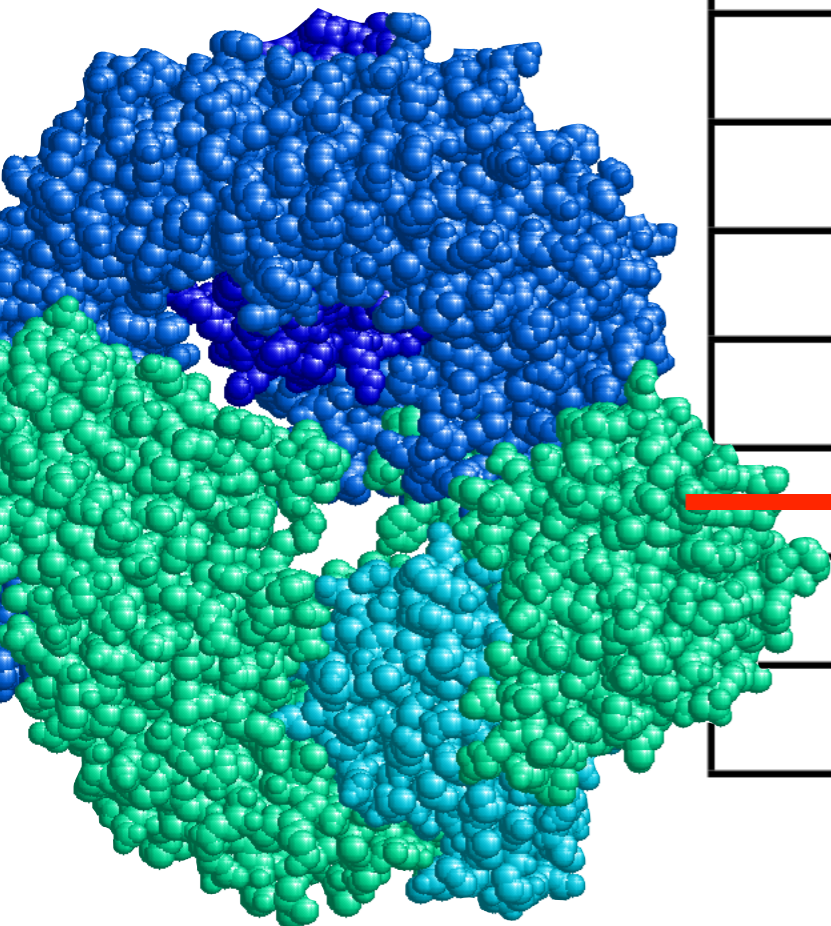
The Sequenome

sequence length	possible sequences
2	4.0E+02
3	8.0E+03
4	1.6E+05
5	3.2E+06
6	6.4E+07
7	1.3E+09
8	2.6E+10
12	4.1E+15
40	1.1E+52
80	1.2E+104
160	1.5E+208
300	2.0E+390
320	2.1E+416
1000	1.1E+1301

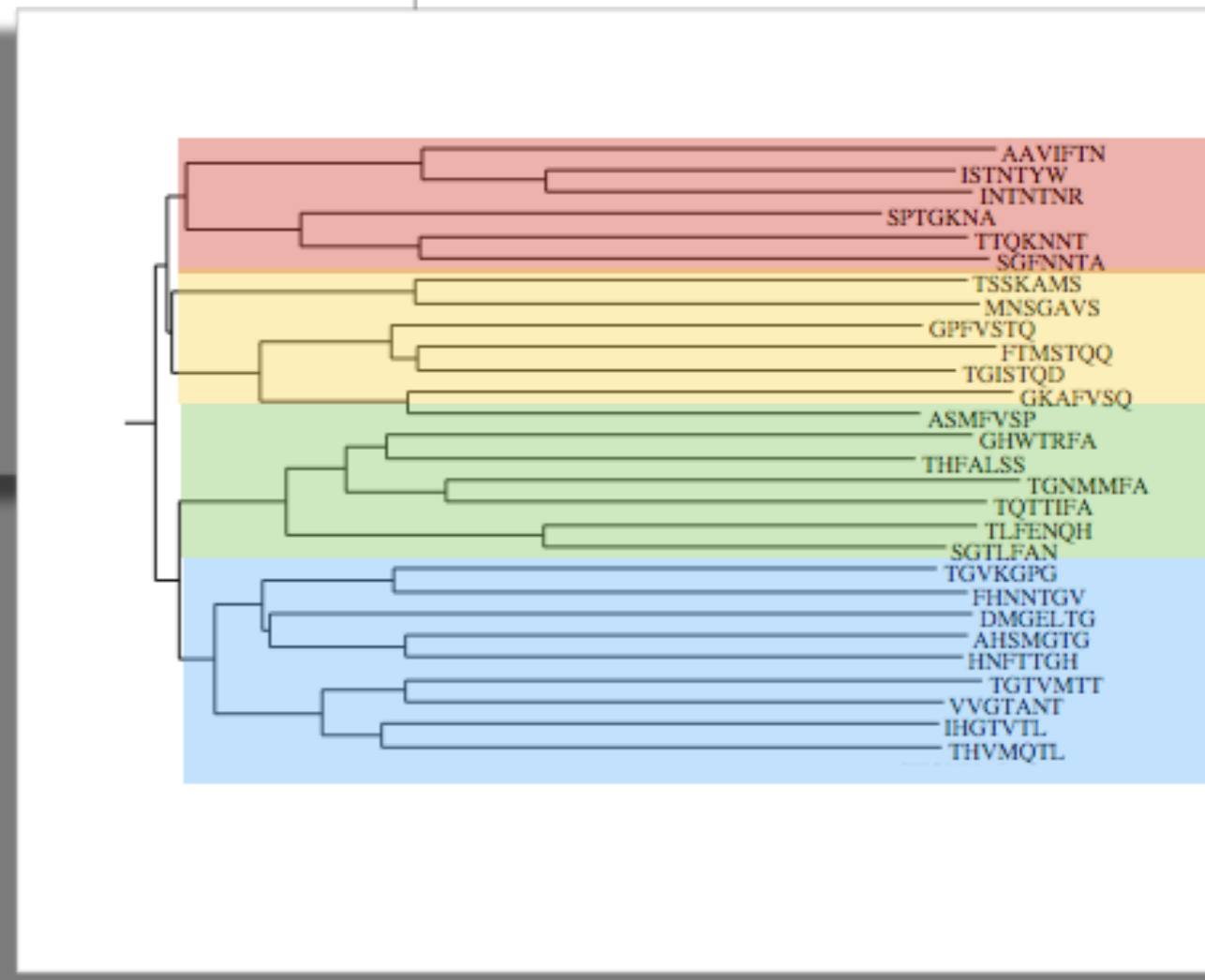
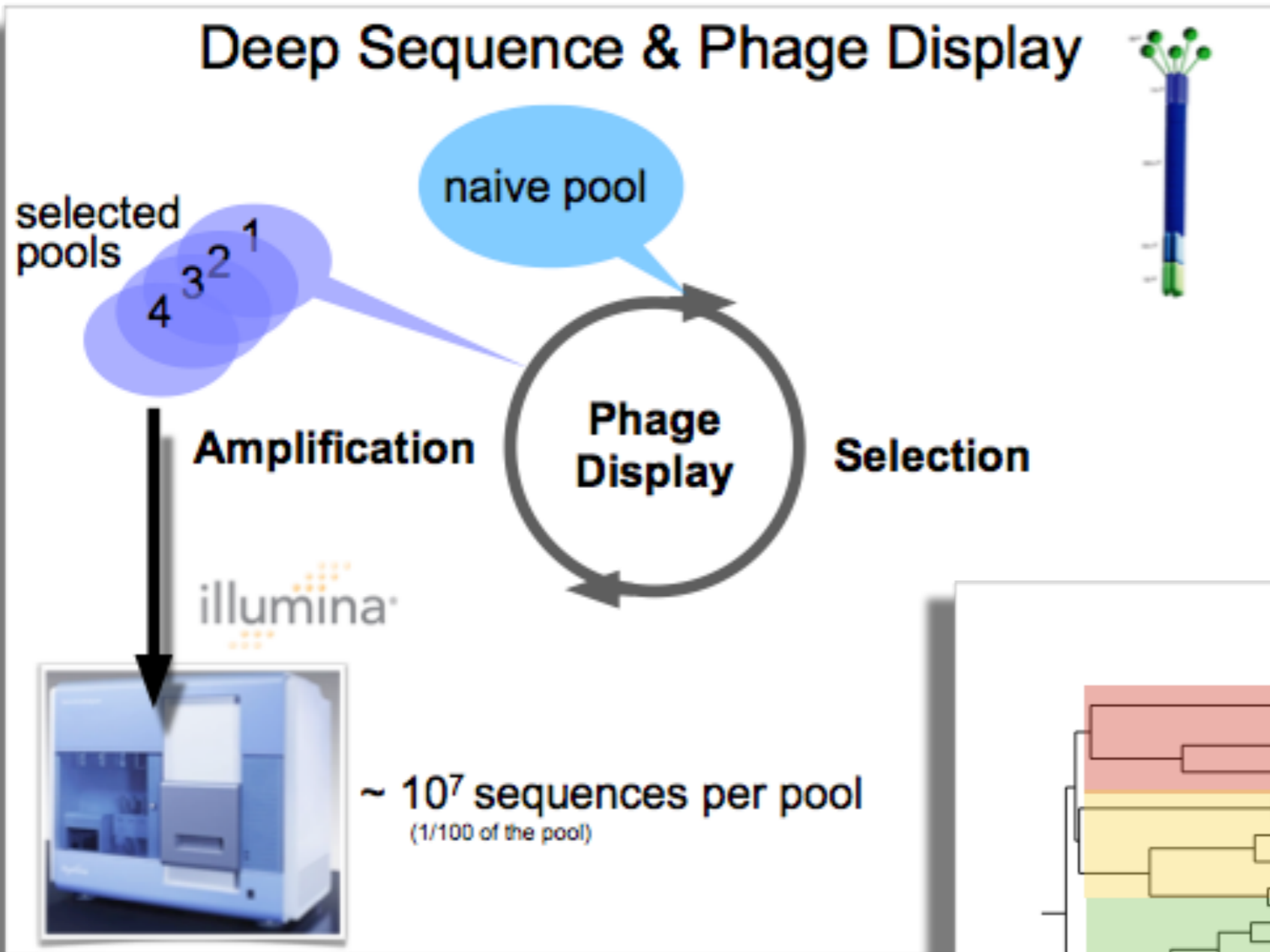


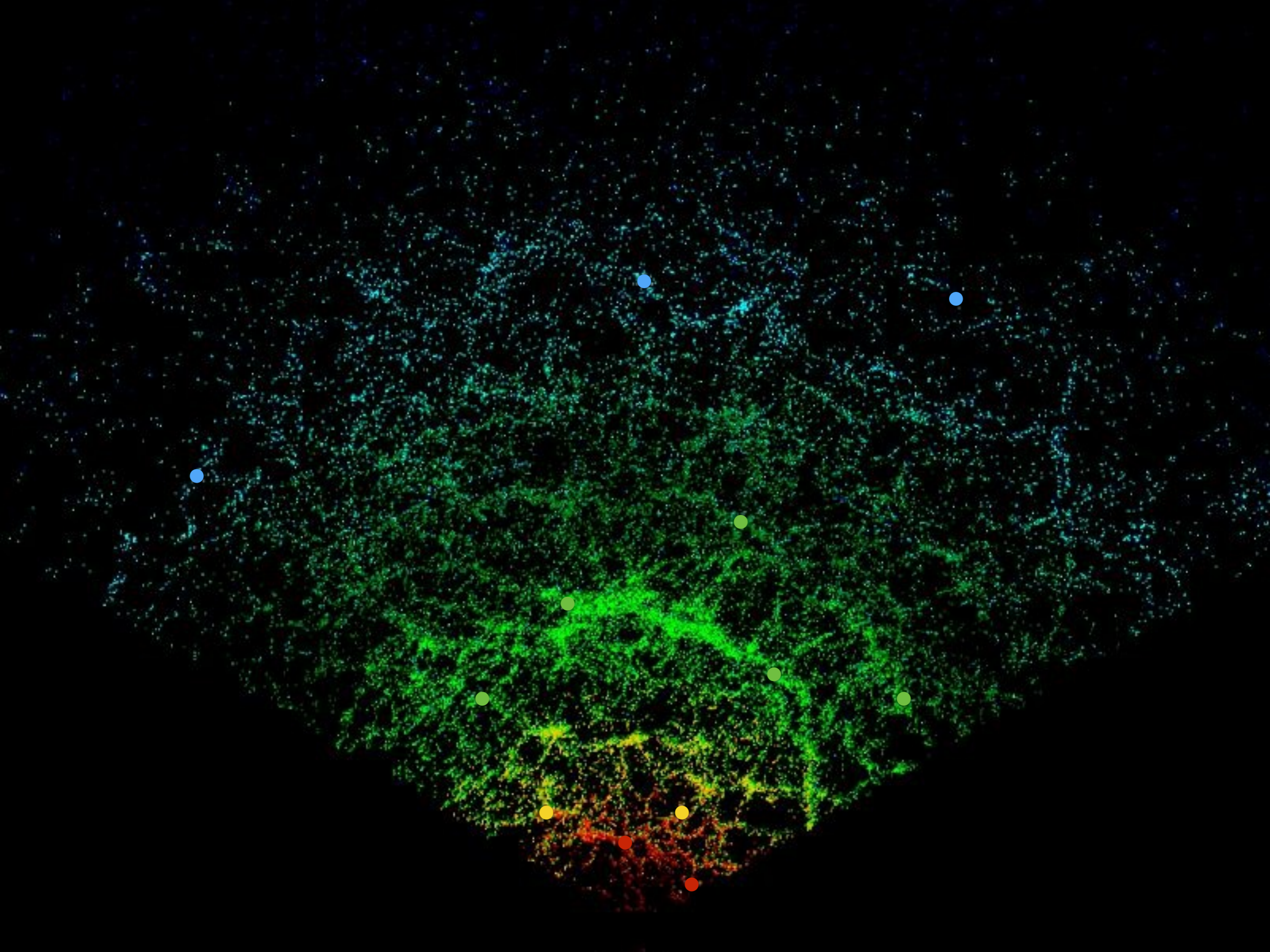
The Sequenome

sequence length	possible sequences
2	4.0E+02
3	8.0E+03
4	1.6E+05
5	3.2E+06
6	6.4E+07
7	1.3E+09
8	2.6E+10
12	4.1E+15
40	1.1E+52
80	1.2E+104
160	1.5E+208
300	2.0E+390
320	2.1E+416
1000	1.1E+1301



The 7-mer Sequenome: 1.28B sequences





The 7-mer Sequenome: 1.28B sequences

Fundamental Questions:

- How many structures are there?
- How are structures distributed ?
- For any given structure, what is the range of kinetics and thermodynamics?
- How do these distributions map onto Silvana's phage-display data (LUMC)?
- Comparisons with PepBank, SAROTOP

	Bioinformatics	Computer Science	Computer Science	Business ICT
Back-ground	<ul style="list-style-type: none"> we have now very large data base of computed structures for all possible 1.28 billion 7-mer peptides, at 3 different temperatures 	<ul style="list-style-type: none"> we have now very large data base of computed structures for all possible 1.28 billion 7-mer peptides, at 3 different temperatures we have now light weight NK model (with Emmerich) 	<ul style="list-style-type: none"> we have now a working SeDEx prototype http://one-34408.sedex-lumc.cloudlet.sara.nl/SeDEx/ 	<ul style="list-style-type: none"> we have now a working SeDEx prototype http://one-34408.sedex-lumc.cloudlet.sara.nl/SeDEx/
Aim of the research	<ul style="list-style-type: none"> First ever look at protein sequence space (many new questions) show application to highly active research area (phage-display) 	<ul style="list-style-type: none"> fit structure data to the NK model of rugged fitness landscapes look for critical transitions relate to complexity classes 	<ul style="list-style-type: none"> upgrade to enterprise-ready SeDEx: a web platform to make <i>in silico</i> protein folding easy and cheap 	<ul style="list-style-type: none"> plan the launch the SeDEx as a commercial platform
Project	<ul style="list-style-type: none"> statistical analysis on 7-mer space correlations with LUMC experimental data correlations with other peptide databases 	<ul style="list-style-type: none"> explore parameter space of NK model to get a good fit to the computed structure landscape for all 7-mers 	<ul style="list-style-type: none"> scale & performance (Nerdalize) add features: <ul style="list-style-type: none"> - search function - platform analytics - licensing / payment tools 	<ul style="list-style-type: none"> develop business cases (market research) develop valorisation plan <ul style="list-style-type: none"> - STW funding / private investment - 'advertising' & platform
Skills	<ul style="list-style-type: none"> R, Matematica, PYTHON Big Data simple and advanced statistics working with multiple peptide databases 	<ul style="list-style-type: none"> analytical CS skills good programming Big Data evolution and optimisation theory (if you like it) 	<ul style="list-style-type: none"> enterprise scale software design website development databases writing APIs 	<ul style="list-style-type: none"> market analysis cost / revenue models negotiation skills working with LUMC & LURIS
Work plan	<ul style="list-style-type: none"> start a.s.a.p. end with short high-profile paper 	<ul style="list-style-type: none"> start a.s.a.p. end with short high-profile paper making a bridge between theory and data 	<ul style="list-style-type: none"> start a.s.a.p. end with upgraded web platform (mock pages where necessary) 	<ul style="list-style-type: none"> start a.s.a.p. end with recommendations for CS project (payment tools, API) end with draft of STW

Knowledge Dynamics: Knowledge representation & reasoning

Problem: Biological Complexity & Data Overload

Knowledge Dynamics: Knowledge representation & reasoning

Problem: Biological Complexity & Data Overload

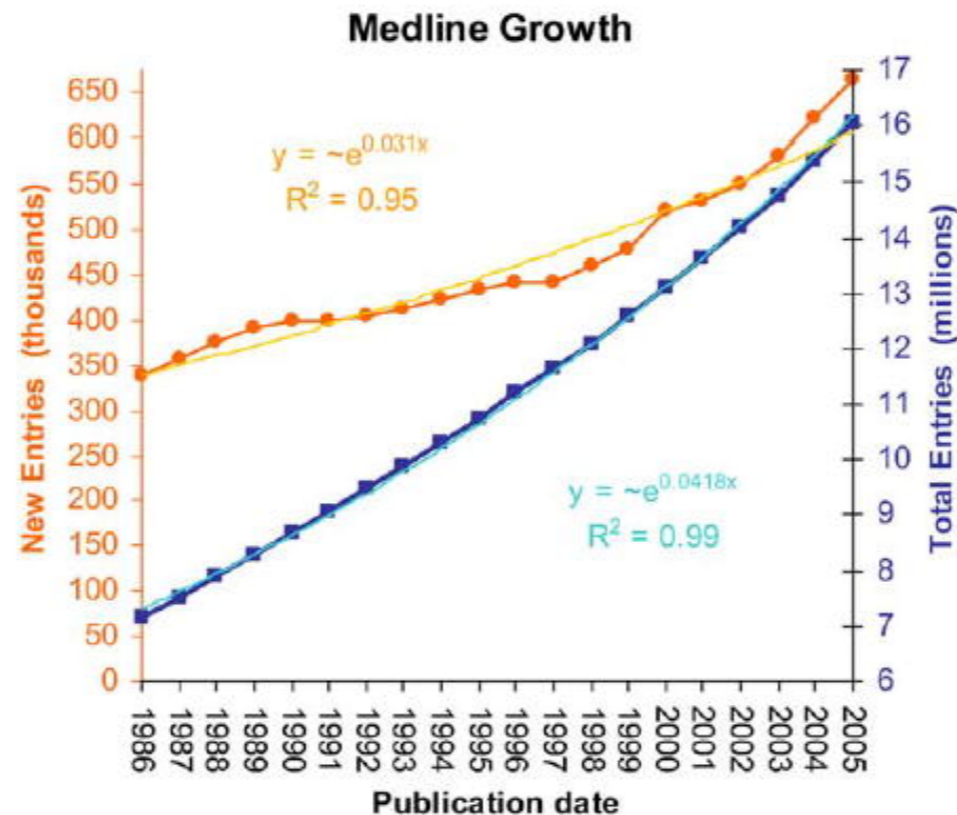


Knowledge Dynamics: Knowledge representation & reasoning

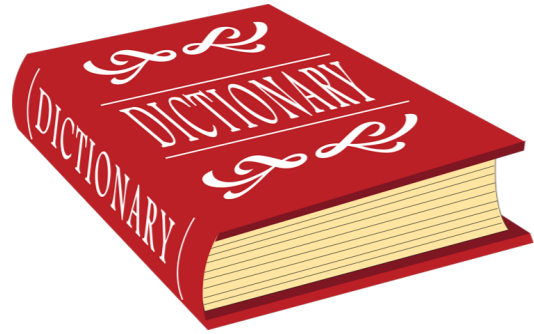
Problem: Biological Complexity & Data Overload



- 21 million references (abstracts) to journal articles
- 1946 - present (January 1980...)
- 5,600 worldwide journals
- More than 700,000 references added in 2013
- 2,000-4,000 references are added each day (Tuesday - Saturday)
- 80 references are added each hour



LWAS



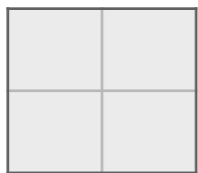
- Dictionary: 687,718 biomedical concepts
- genes, diseases, symptoms, biological processes
- disambiguation
- UMLS+Entrez Gene+OMIM+UniProt



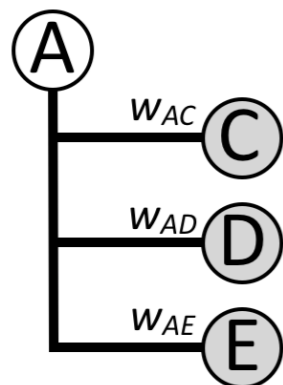
- Corpus: 17M PubMed abstracts from January 1980



- Index all concepts per abstract

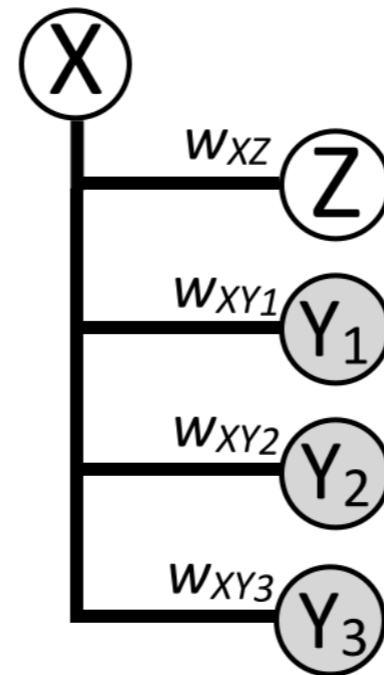
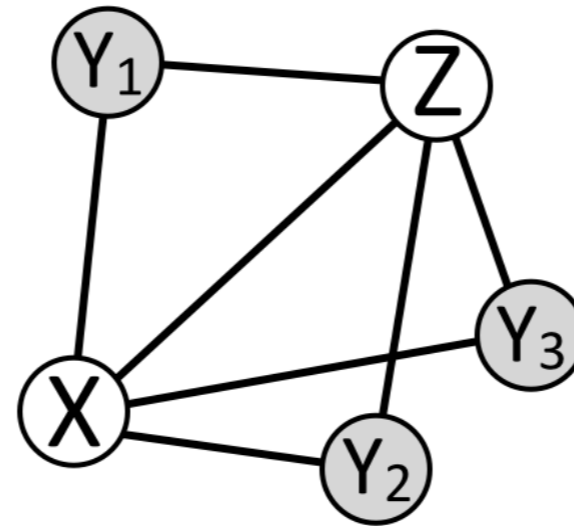


- Compute co-occurrence frequencies of concept pairs

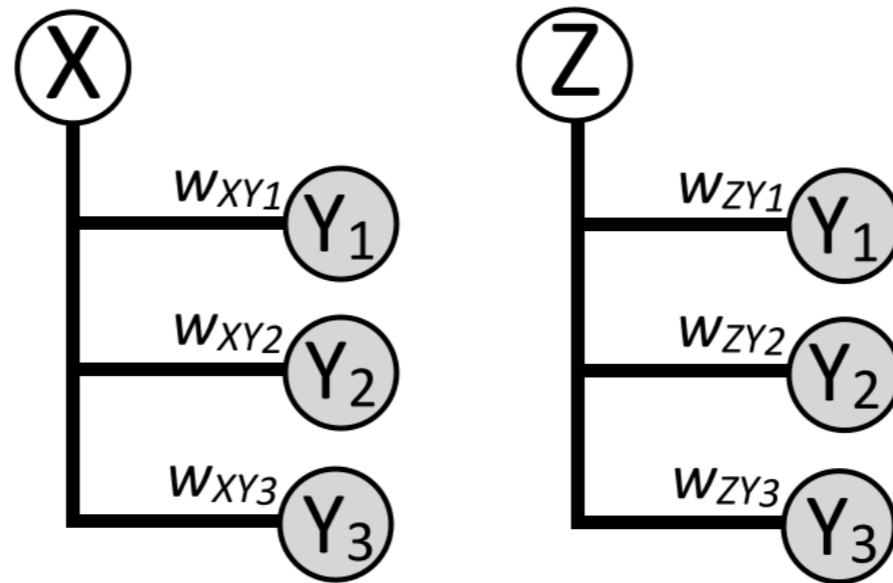
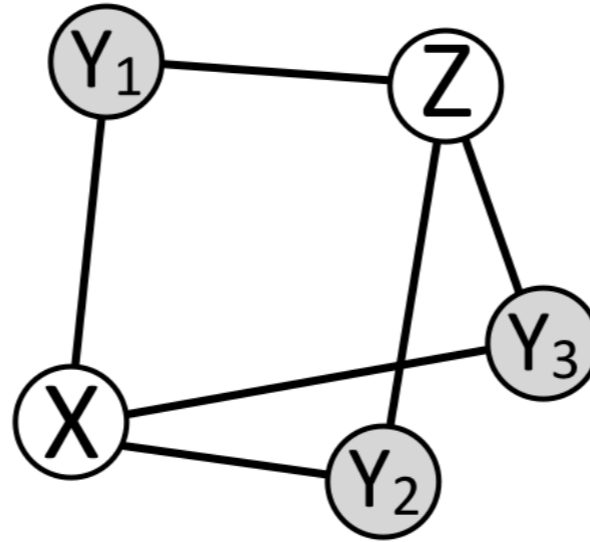


- Build concept profiles
- For any concept, a list of all other concepts
- Weighted by UC of the co-occurrence frequencies
- Minimally 5 abstracts, maximum 140,000 concepts
- Expose associations independently of documents

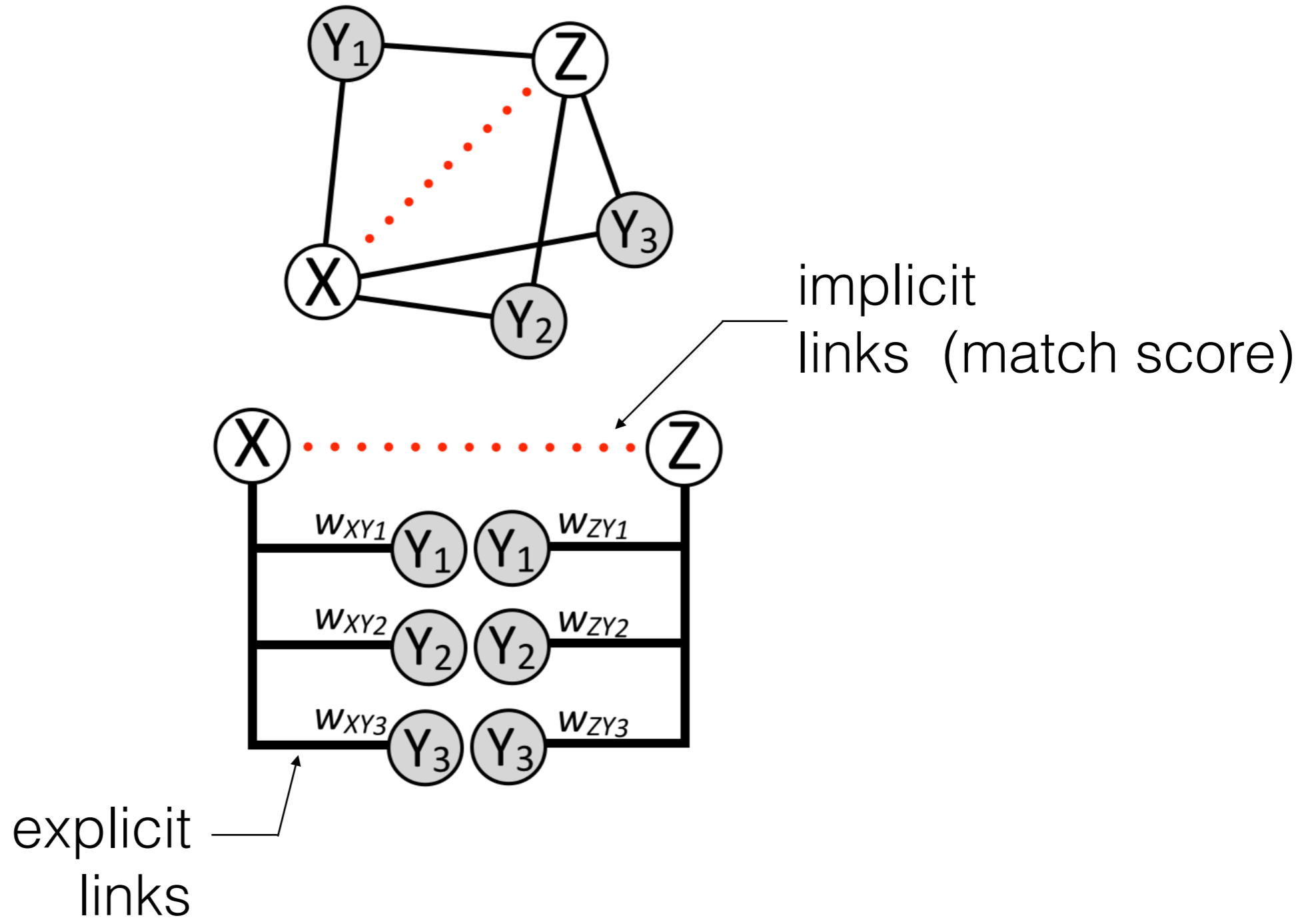
Knowledge Dynamics: Knowledge representation & reasoning



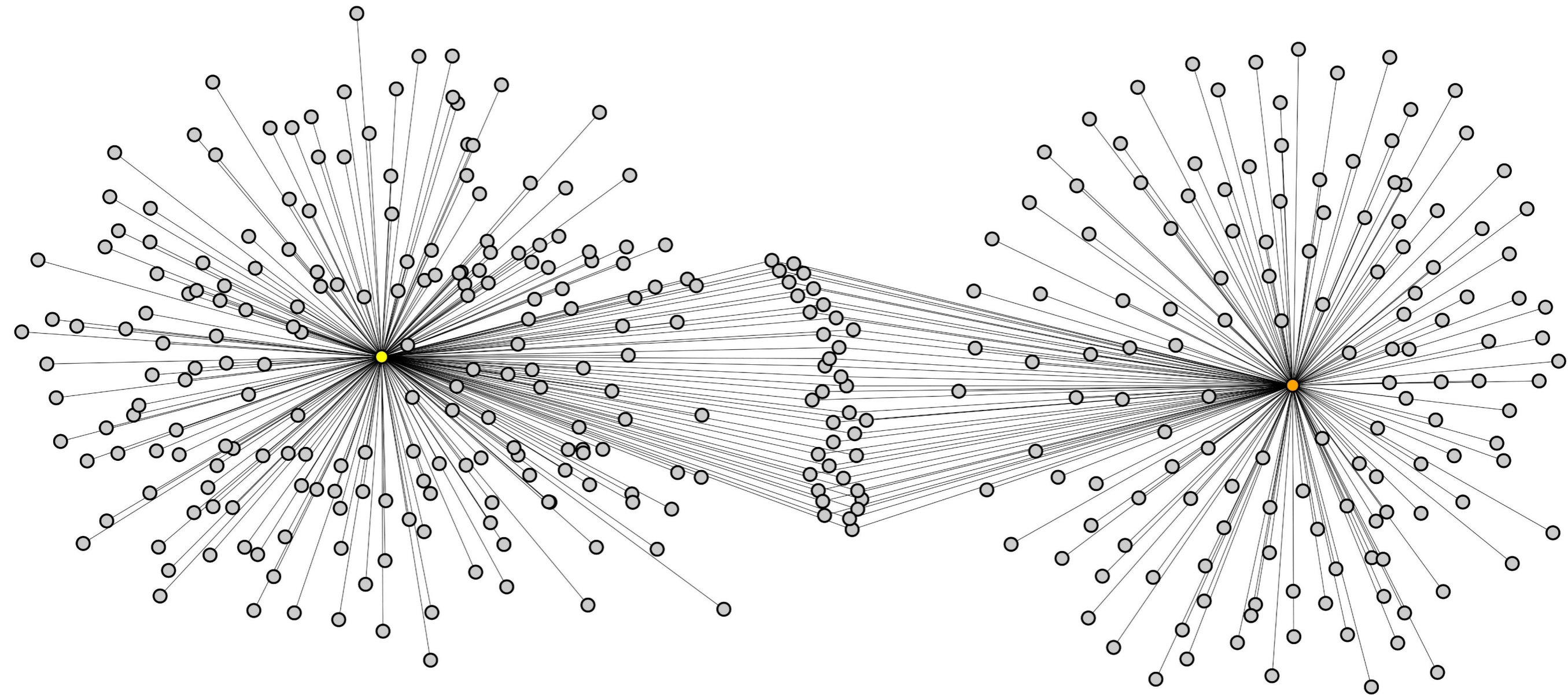
Knowledge Dynamics: Knowledge representation & reasoning



Knowledge Dynamics: Knowledge representation & reasoning



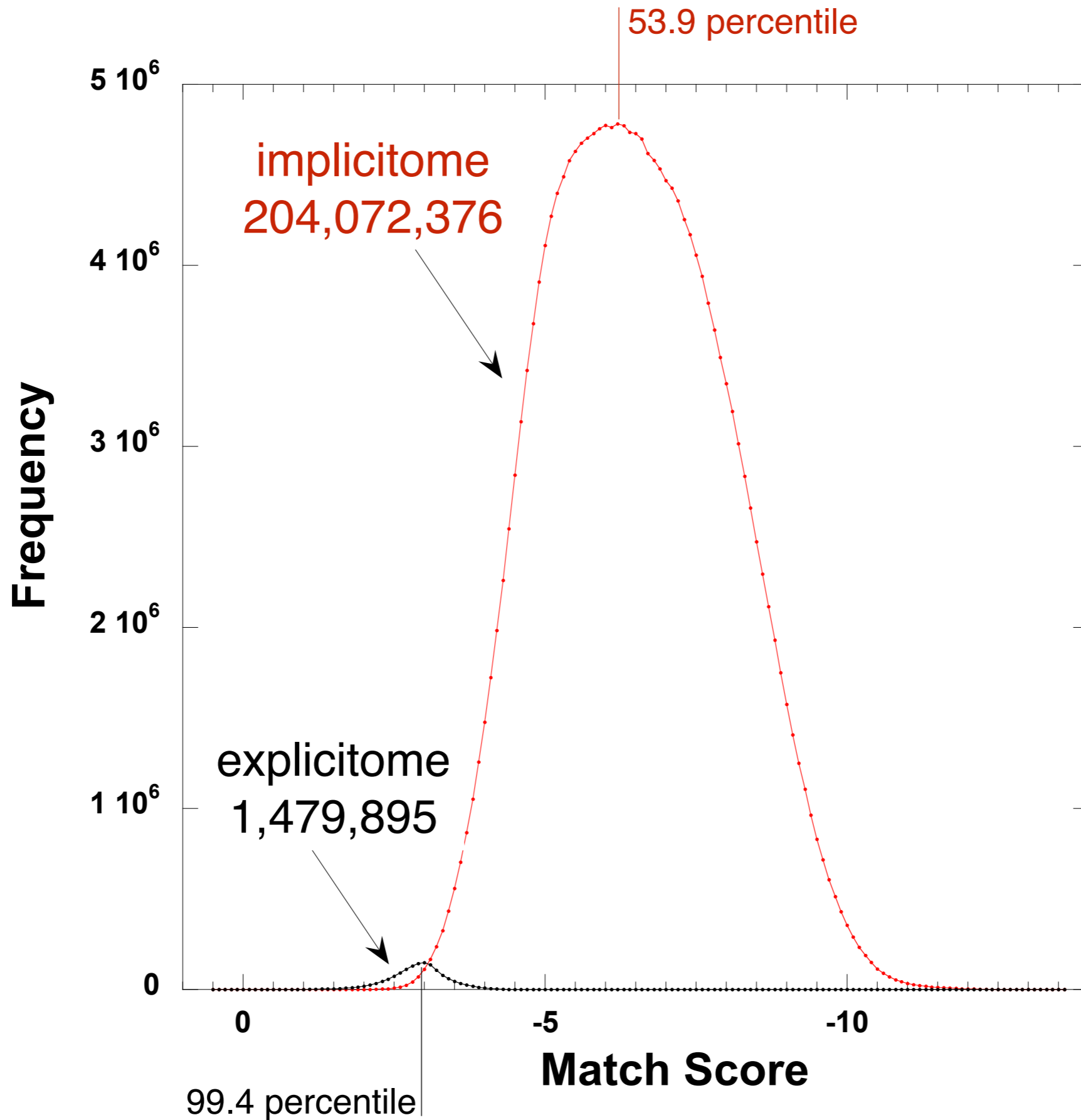
Knowledge Dynamics: Knowledge representation & reasoning



CWH43

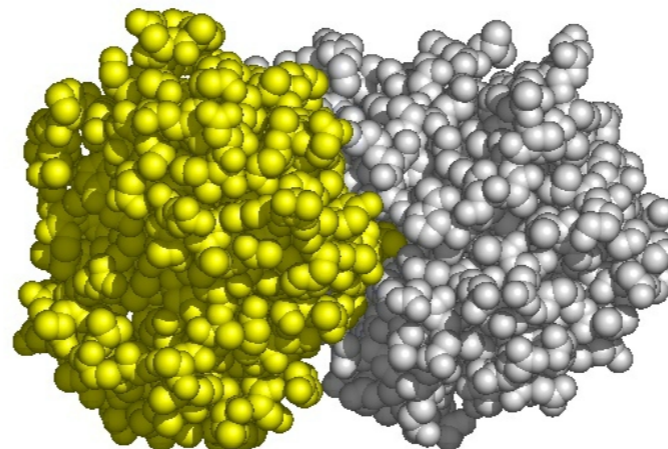
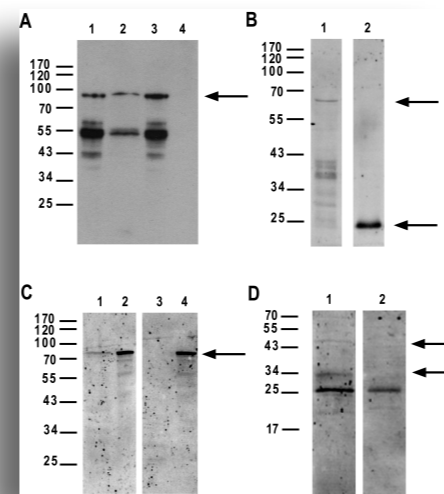
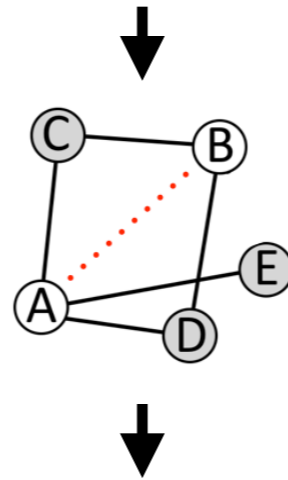
Hyperphosphatesia,
Mental Retardation

Knowledge Dynamics: Knowledge representation & reasoning



Example: *Physical interaction between CAPN3 & PARVB*

PubMed



Example: *Physical interaction between CAPN3 & PARVB*

PLOS ONE | OPEN ACCESS | PEER-REVIEWED RESEARCH ARTICLE

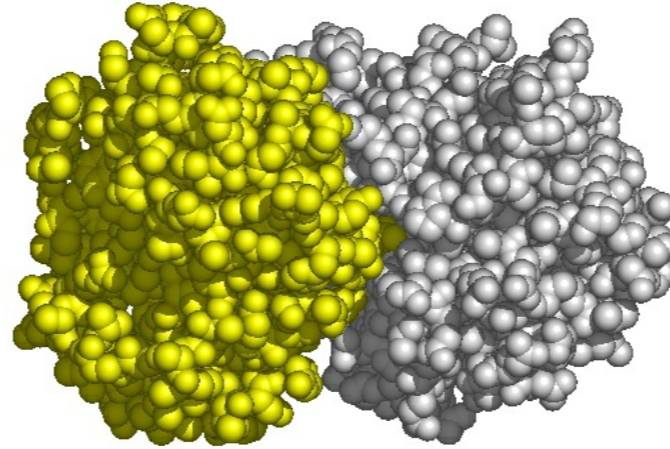
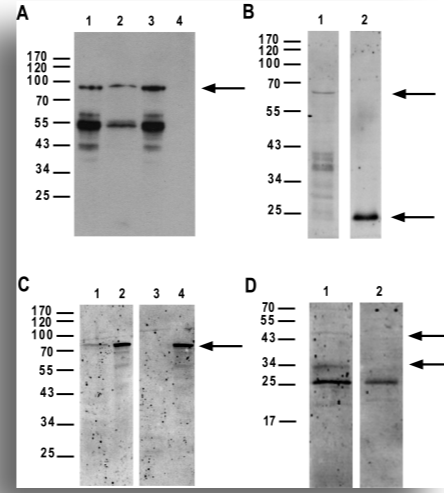
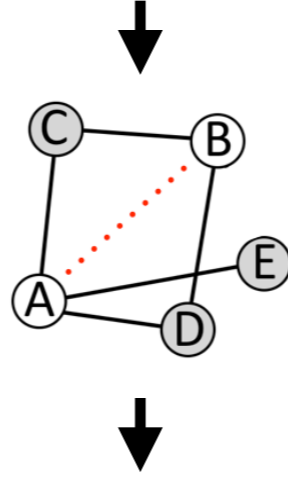
advanced search

Search 6,236 VIEWS 15 CITATIONS 60 SAVES 1 SHARE

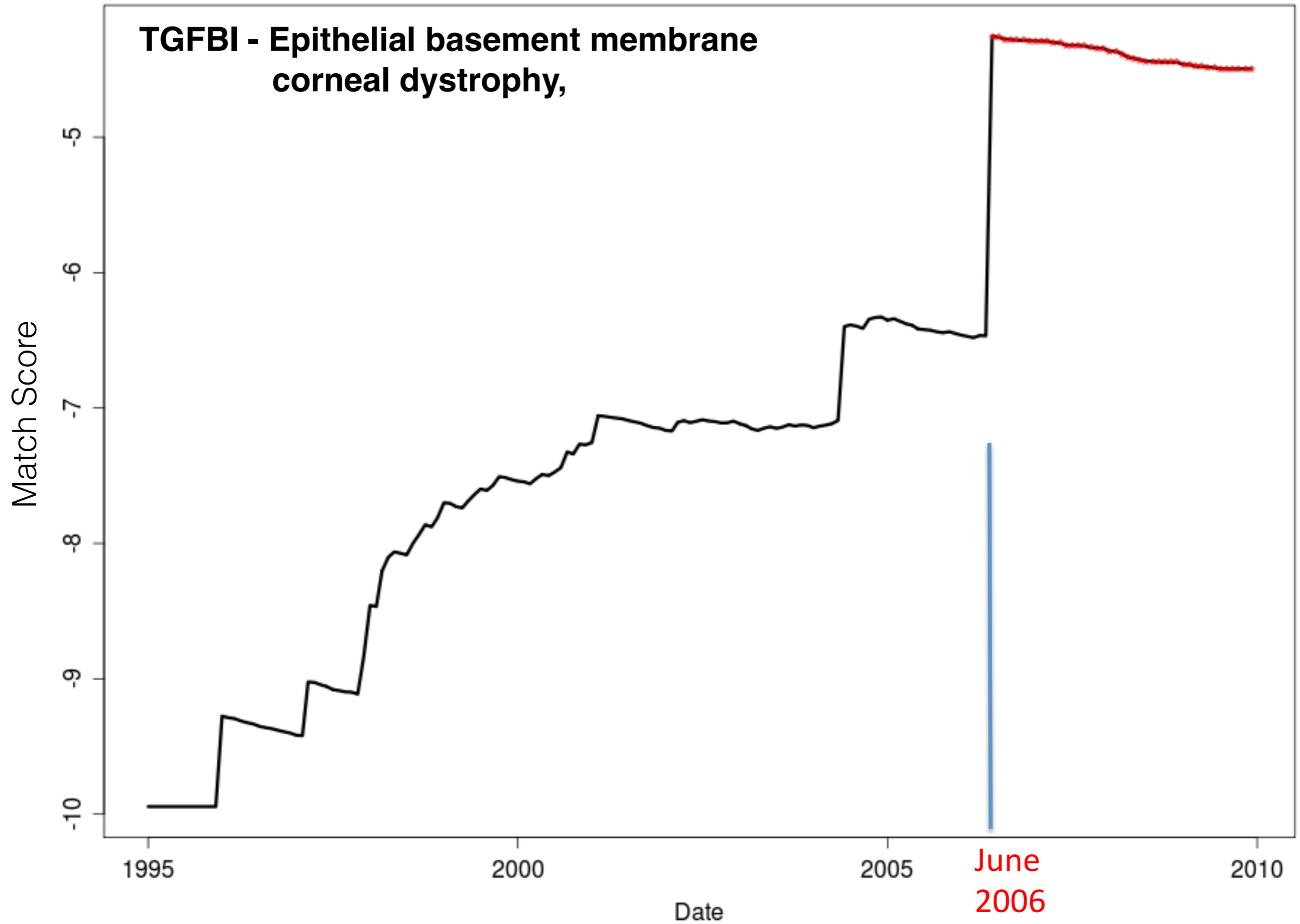
Novel Protein-Protein Interactions Inferred from Literature Context

Herman H. H. B. M. van Haagen, Peter A. C. 't Hoen, Alessandro Botelho Bovo, Antoine de Morrée, Erik M. van Mulligen, Christine Chichester, Jan A. Kors, Johan T. den Dunnen, Gert-Jan B. van Ommen, Silvére M. van der Maarel, Vinicius Medina Kern, Barend Mons, Martijn J. Schuermie

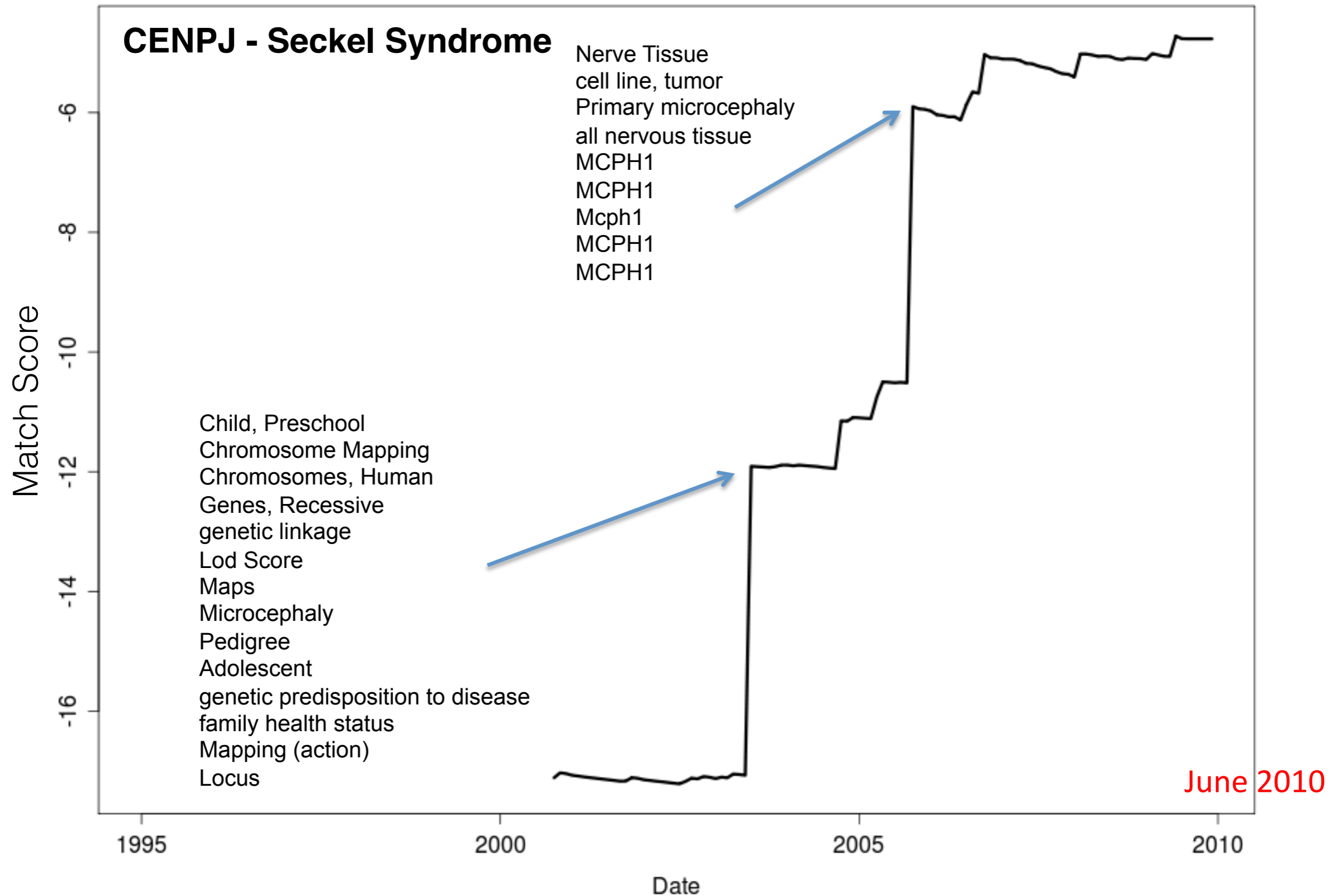
Published: November 18, 2009 • DOI: 10.1371/journal.pone.0007894



Knowledge Dynamics: Knowledge representation & reasoning



Knowledge Dynamics: Knowledge representation & reasoning

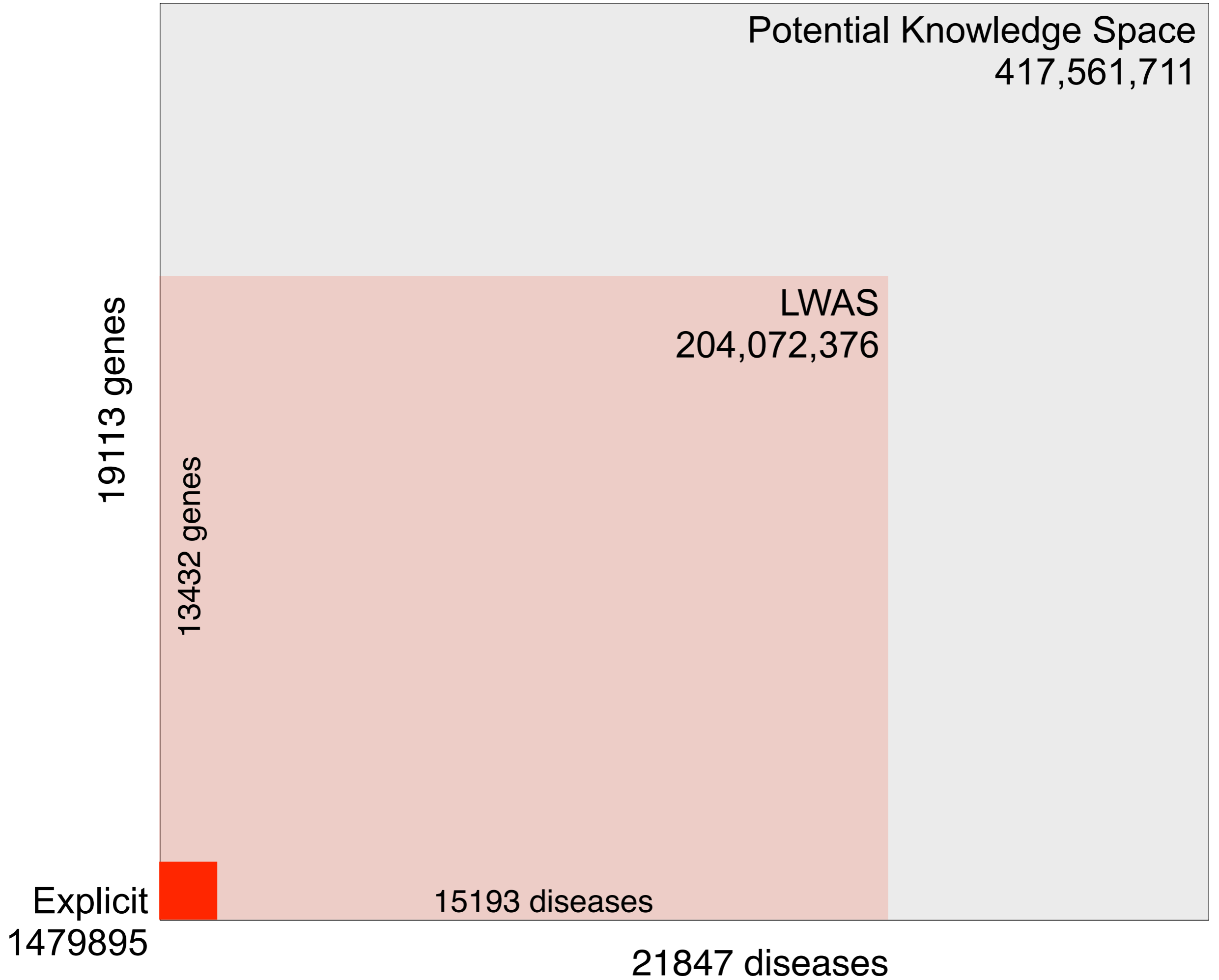


Knowledge Dynamics: Knowledge representation & reasoning



Look at trends in variables in concept profiles over time:

- rate of growth of concept profile
- distribution of weights
- distribution of semantic type
- distribution of generic / specific concepts
- distribution of implicit / explicit concepts
- distribution of network properties (degree, entropy, etc)



	Concept Profiles	Trends
Background	<ul style="list-style-type: none"> we have now 204 million PubMed gene-disease concept profiles 	<ul style="list-style-type: none"> we have now 204 million PubMed gene-disease concept profiles
Aim of the research	<ul style="list-style-type: none"> Probe structure of concept profiles for predictors of landmark discovery 	<ul style="list-style-type: none"> map growth of concept web over time
Project	<ul style="list-style-type: none"> 'filter' gene and disease concept profiles and measure performance in predicting novel gene-disease associations 	<ul style="list-style-type: none"> compute gene-disease associations at regular time points since 1980 get network analytics on growth dynamics of concept web make beautiful animated gifs
Skills	<ul style="list-style-type: none"> R and other scripting simple statistics 	<ul style="list-style-type: none"> running our pipeline to generate time-delimited concept profiles R and other scripting scientific visualisation and animation is important
Work plan	<ul style="list-style-type: none"> start a.s.a.p end with short high-profile paper help drafting Horizon 2020 proposal (March 31) 	<ul style="list-style-type: none"> simple st start a.s.a.p end with short high-profile paper help drafting Horizon 2020 proposal (March 31)