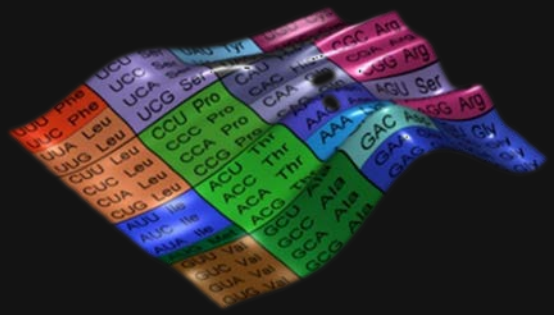
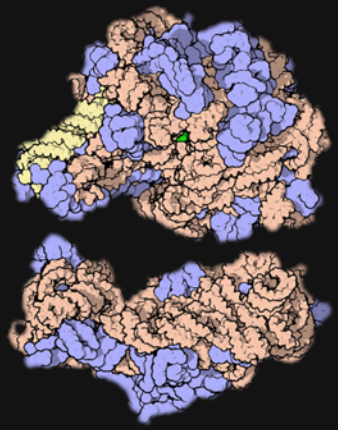
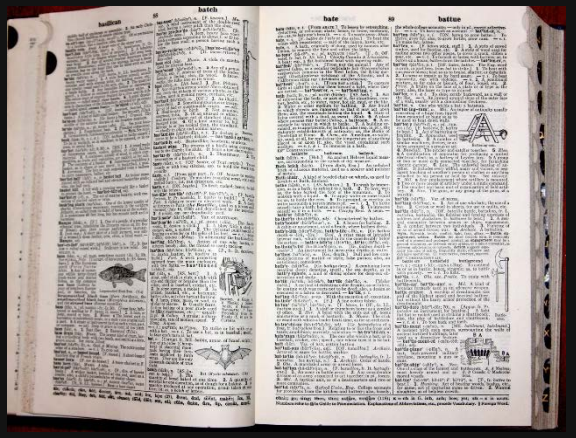
Structures of living information systems







Nov 2013

I. Generic properties of molecular codes?

- Molecular code = map relating two sets of molecules.
- Spaces defined by similarity of molecules (size, polarity etc.)

64 codons

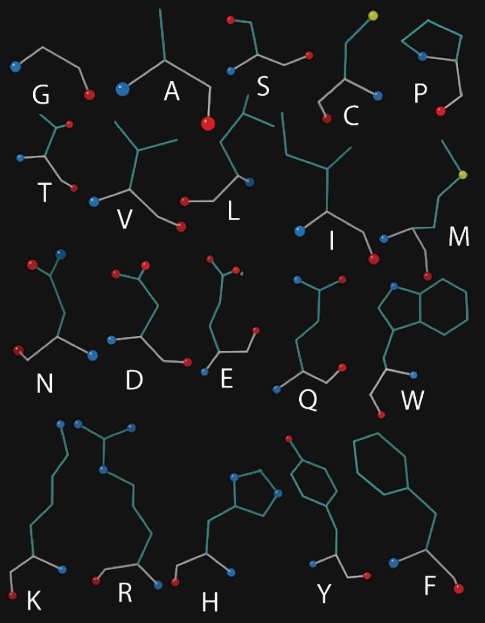
TTT TCT TAT TGT
TTT TCC TAC TGC
TTA TCA TAA TGA
TTG TCG TAG TGG
CTT CCT CAT CGU
CTC CCC CAC CGC
CTA CCA CAA CGA
ATT ACT AAT AGT
ATC ACC AAC AGC
ATA ACA AAA AGA
ATG ACG AAG AGG
GTT GCT GAT GGT
GTC GCC GAC GGC
GTA GCA GAA GGA
GTG GCG GAG GGG

$\text{codon} = b_1 b_2 b_3,\ b_i \in \{\text{A, T, G, C}\}.$

$\phi(\text{codon}) \rightarrow \text{amino-acid.}$

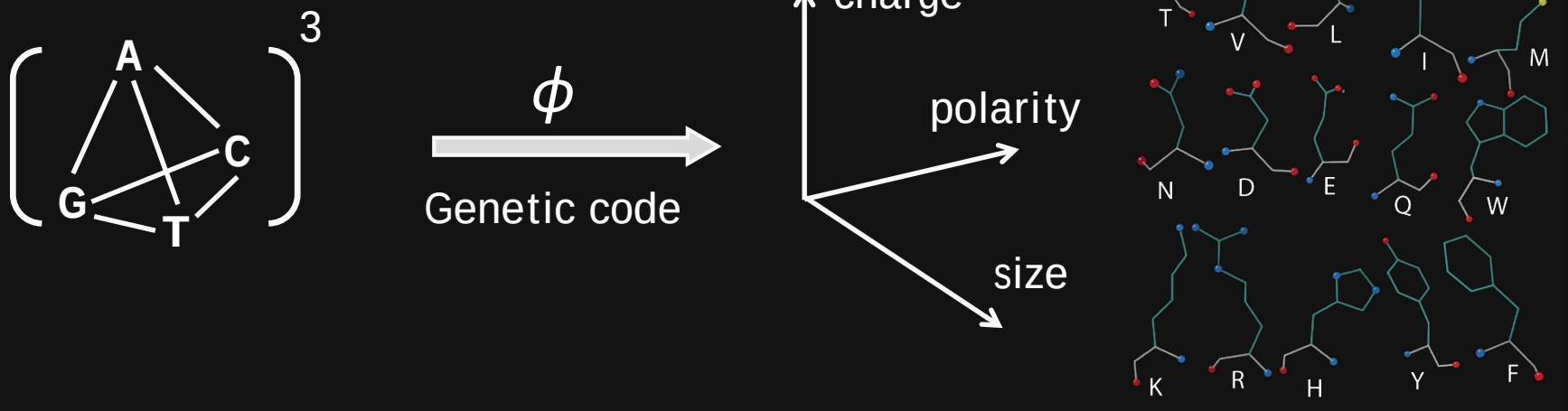
Genetic Code

20 amino-acids



Genetic code maps between DNA and protein languages

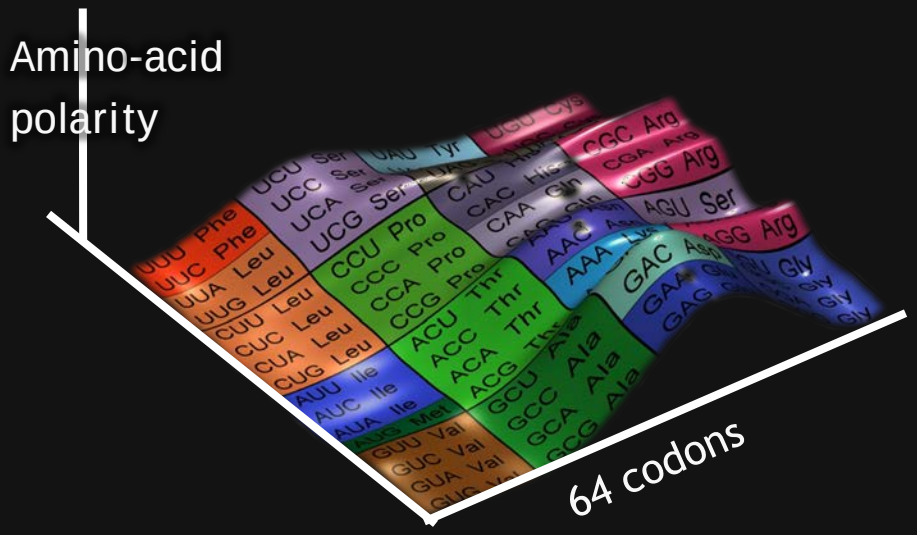
- Code embeds codon-graph (Hamming) into space of amino-acids (“digital to analog”).



- **Translation machinery** facilitates the map.

The genetic code is a smooth mapping

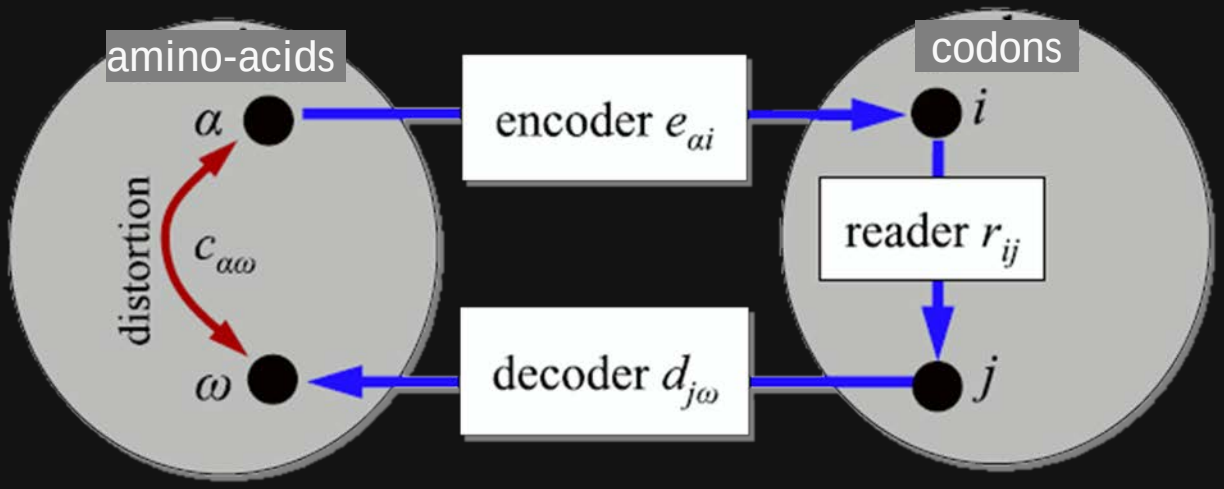
UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA TER	UGA TER
UUG Leu	UCG Ser	UAG TER	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly



- Degenerate (20 out of 64) .
- Compactness of amino-acid regions.
- **Smooth** (similar "color" of neighbors).

Generic properties of molecular codes?

Fitter codes have minimal distortion

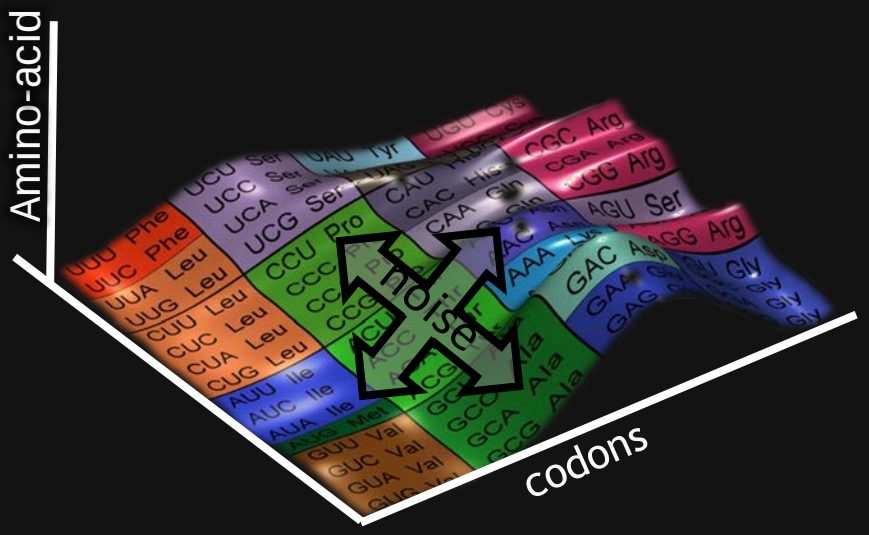


$$\mathrm{D} = \langle c_{\alpha\omega} \rangle = \sum_{paths} p_{path} c_{\alpha\omega} = \mathrm{Tr}\left(e \cdot r \cdot d \cdot c\right)$$

- Distortion of noisy channel, ***D*** = average distortion of AA.
- ***r*** defines structure of codon space.
- ***c*** defines structure of amino-acid space.

TT, *J Theo Bio* 2008, *PNAS* 2008, *Phys Life Rev* 2010

Smooth codes minimize distortion



- Noise confuses close codons.
- Smooth code:

  close codons = close amino-acids.

  → minimal distortion.

- Optimal code balances contradicting needs for **smoothness** and **diversity**.

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA TER	UGA TER
UUG Leu	UCG Ser	UAG TER	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

Max smoothness

Min diversity

Min smoothness

Max diversity

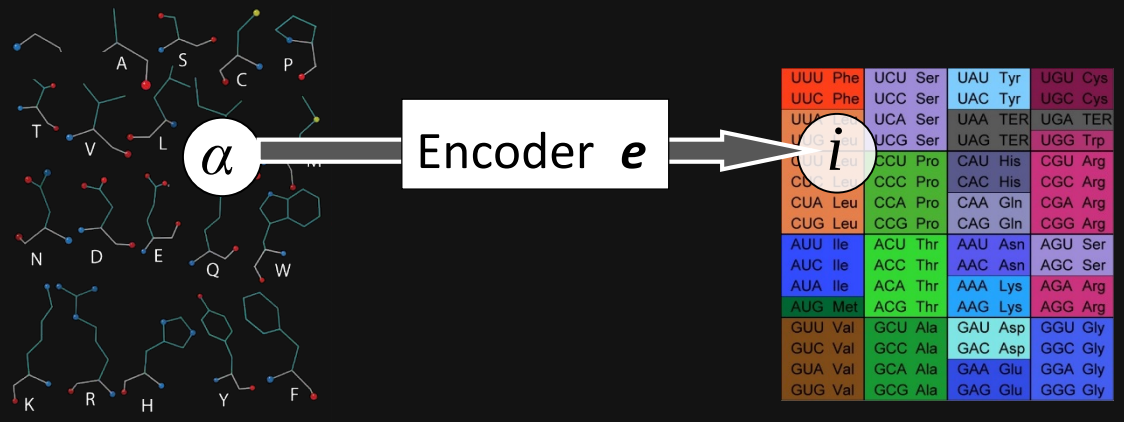
1 — 20 — 64

amino-acids

Channel rate is code's cost

- Diverse codes require high **specificity** = high binding energies ε.
- Cost ~ average $\langle \varepsilon \rangle$.
- Binding prob. ~ Boltzmann: $E \sim e^{\varepsilon/T}$.

$$I \sim \sum_{\alpha,i} e_{\alpha i} \ln e_{\alpha i} \sim \langle \varepsilon_{\alpha i} \rangle_e$$



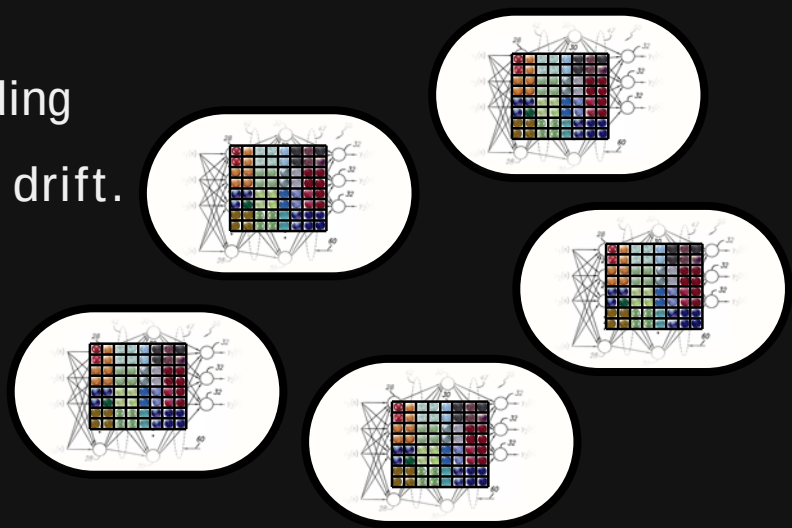
- Cost ***I*** = **Channel Rate** (bits/message)

Code's fitness combines rate and distortion of map

Fitness = *Gain* x *Distortion* + *Rate* $\qquad H = \beta D + I$

- ***Gain*** β increases with organism complexity and environment richness.
- Fitness ***H*** ~ "free energy" with "temperature" $1/\beta$.
- Evolution varies β.

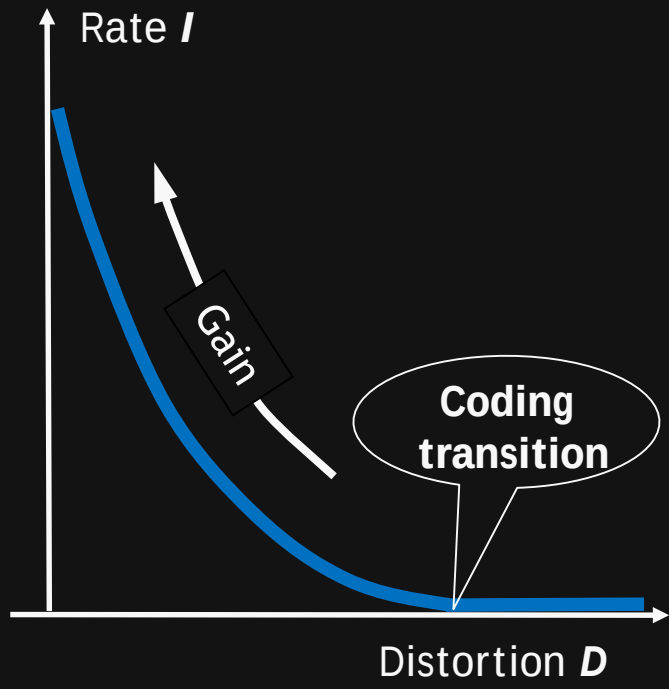
- Population of self-replicators evolving according to code fitness *H*: mutation, selection, random drift.

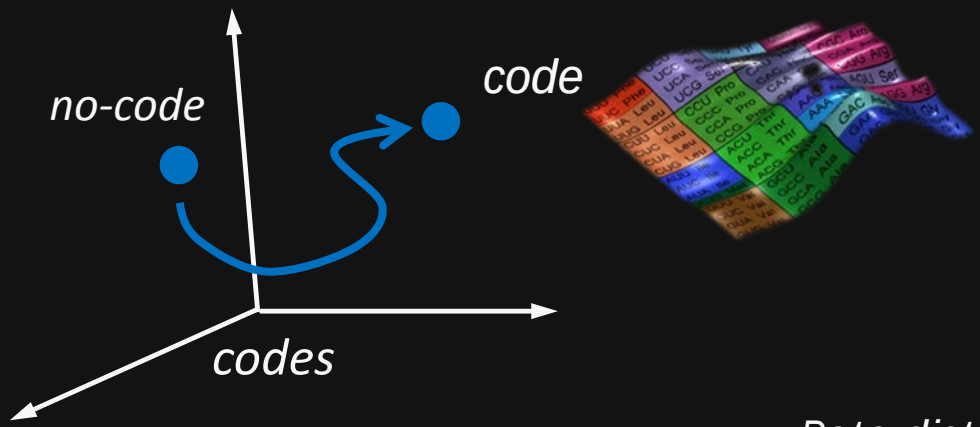


Code emerges at a critical coding transition

- Low gain β : Cost too high
  → no specificity → **no code**.
- **Code emerges** when β increases:
  channel starts to convey information ($I \neq 0$).

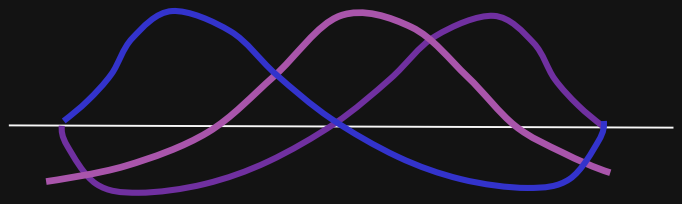
- Continuous phase transition.
- Emergent code is smooth, low mode of $\boldsymbol{r}$.





Rate-distortion theory (Shannon 1956)

Emergent code is a smooth mode of error-Laplacian

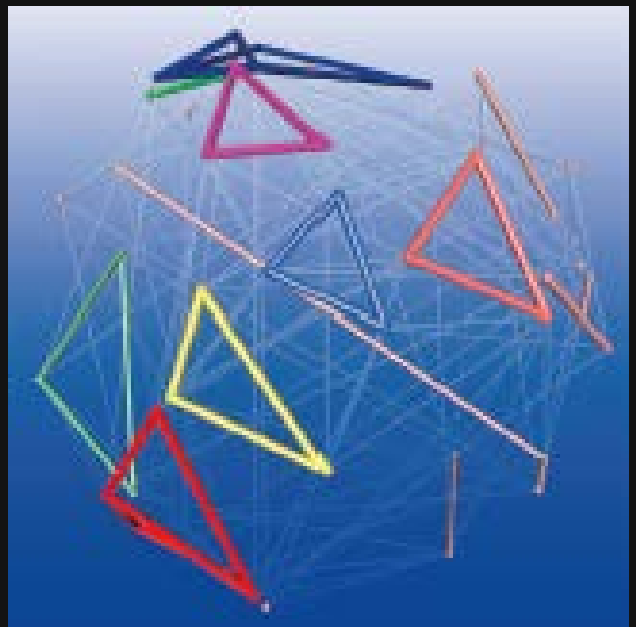


- Lowest excited modes of graph-Laplacian *r* .
- Single maximum for lowest excited modes (Courant).
- Every mode corresponds to amino-acid :

 # low modes = # amino-acids.

→ single contiguous domain for each amino-acid.

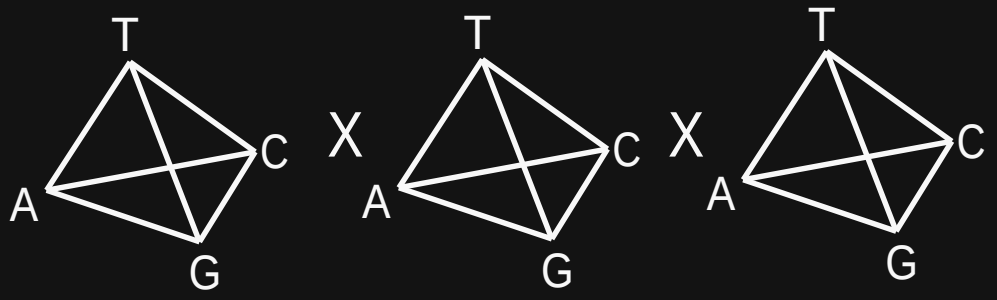
→ **Smoothness**.



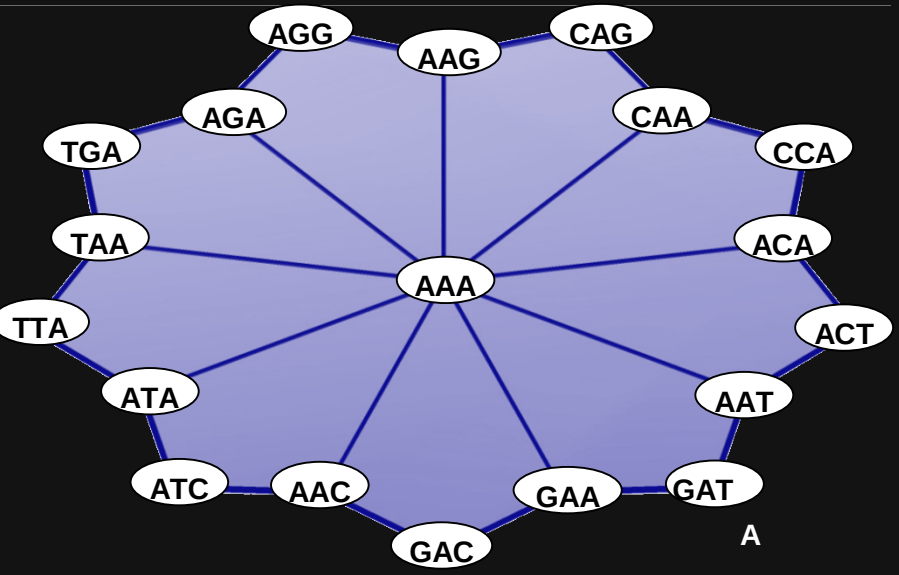
J-P Eckmann, PNAS 2008

Probable errors define the graph and the topology of the genetic code

- Codon graph = codon vertices + 1-letter difference edges (mutations).



$K_4 \times K_4 \times K_4$



- Non-planar graph (many crossings).
- Genus γ = # holes of embedding manifold.
- Graph is holey : embedded in $\gamma = 41$

  (lower limit is $\gamma = 25$)

Coloring number limits number of amino-acids

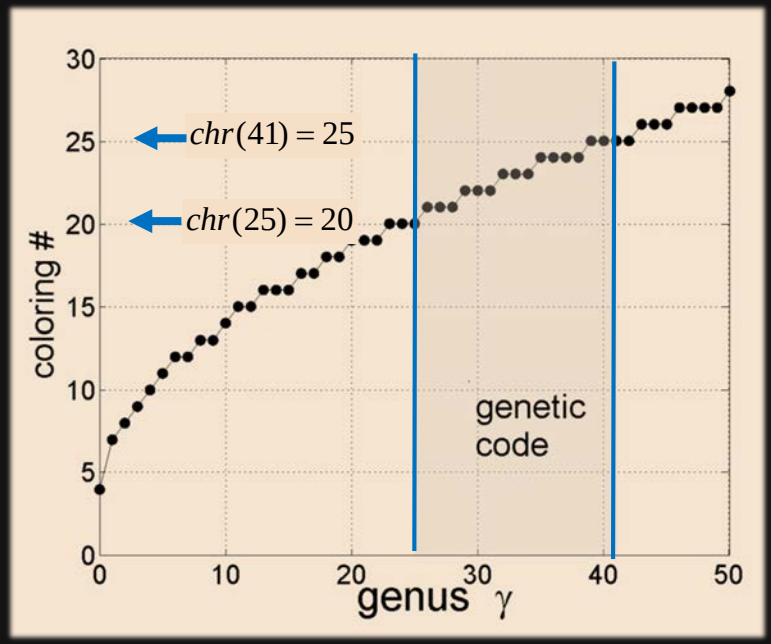
- Minimal # colors to for a map =

  Coloring number, topological invariant (function of genus):

$$chr(\gamma) = \left\lfloor \frac{1}{2}\left(7 + \sqrt{1+48\gamma}\right) \right\rfloor .$$

(Ringel & Youngs 1968)

$$\max(\#\ \text{amino-acids}) = chr(\gamma)$$



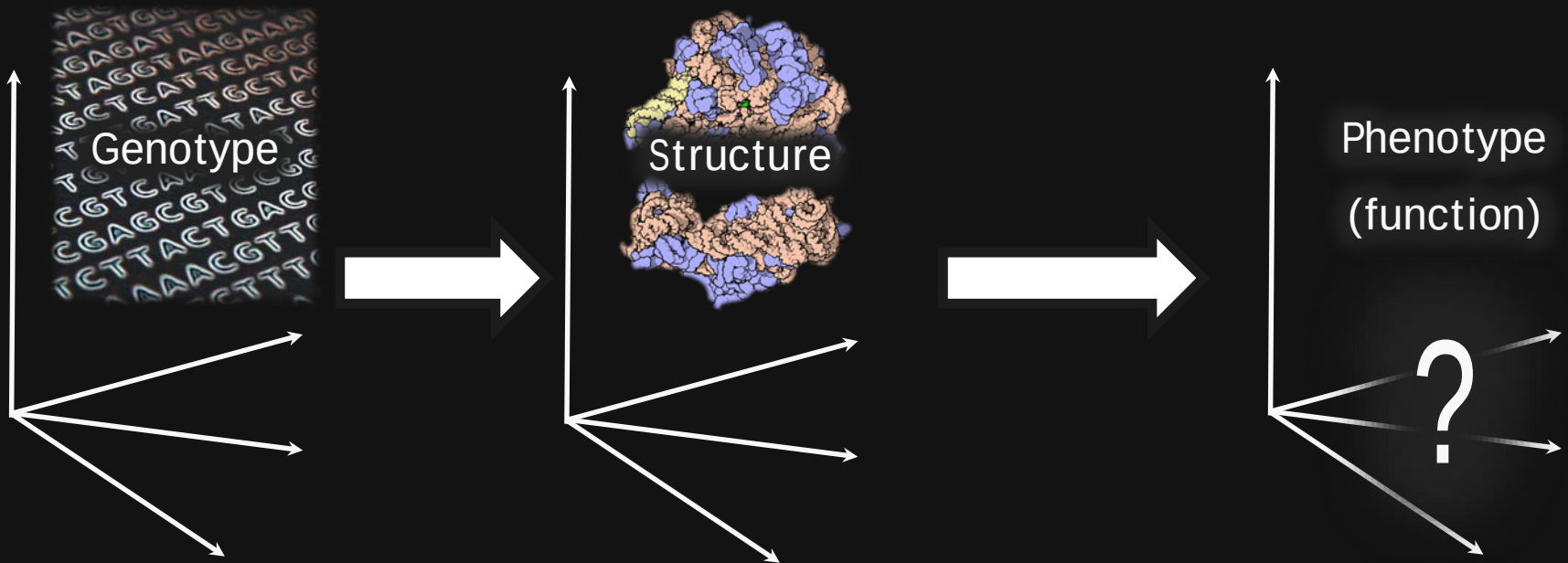
- From Courant 's theorem + "convexity" (tightness).
- Genetic code: γ = 25-41 → coloring number = 20-25 amino-acids

The genetic code coevolves with accuracy

- A path for evolution of codes: from early codes with higher codon degeneracy and fewer amino acids to lower degeneracy codes with more amino acids.

1st	2nd	3rd	γ	chr #
1	4	1	0	**4**
2	4	1	1	**7**
4	4	1	5	**11**
4	4	2	13	**16**
4	4	3	25	**20**
4	4	4	41	**25**

II. What is the dimension of phenotype space?

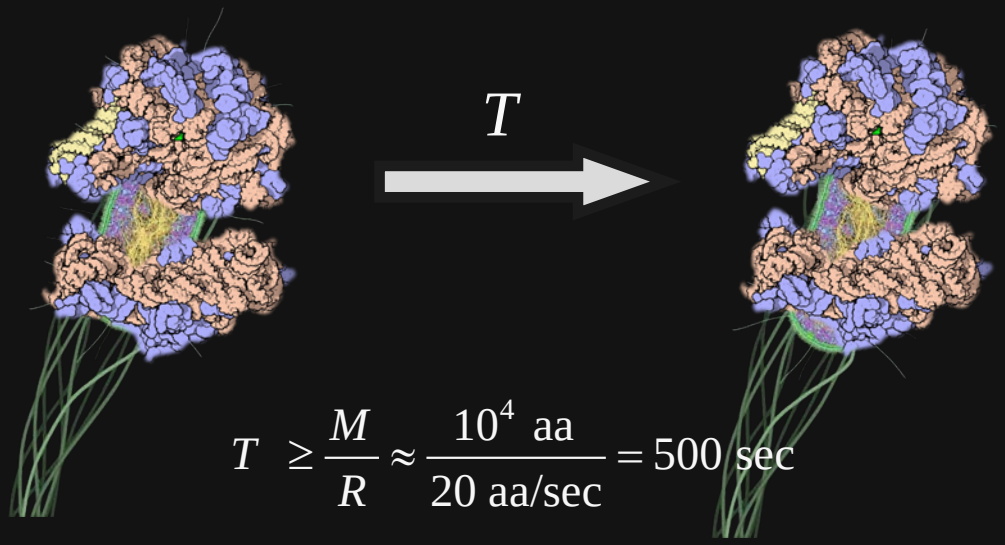


Dimension = # parameters ~ 10^4

What are the relevant degrees of freedom?

Savir & TT, *PNAS* 2010, *Mol Cell* 2010, *Cell* 2013

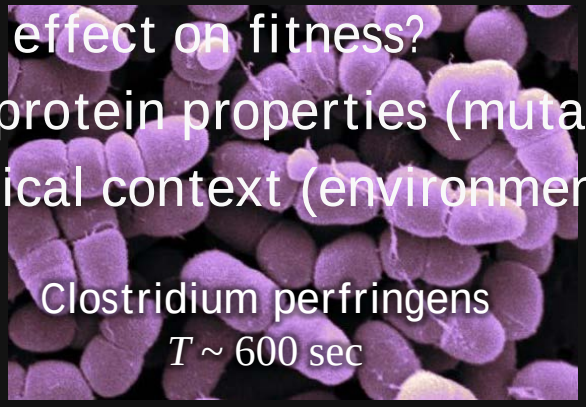
Ribosome limits self-replication rate

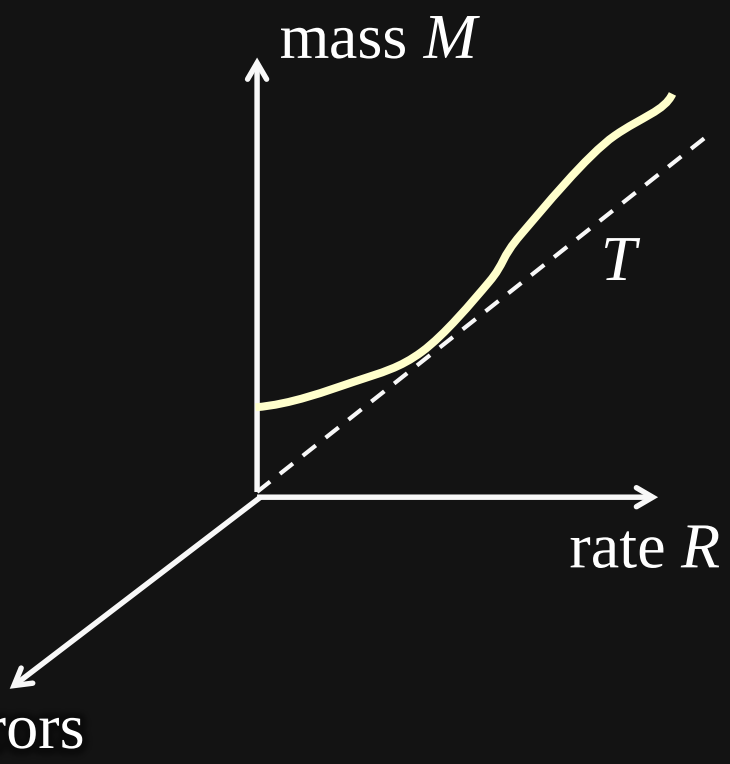


$$T \geq \frac{M}{R} \approx \frac{10^4 \text{ aa}}{20 \text{ aa/sec}} = 500 \text{ sec}$$

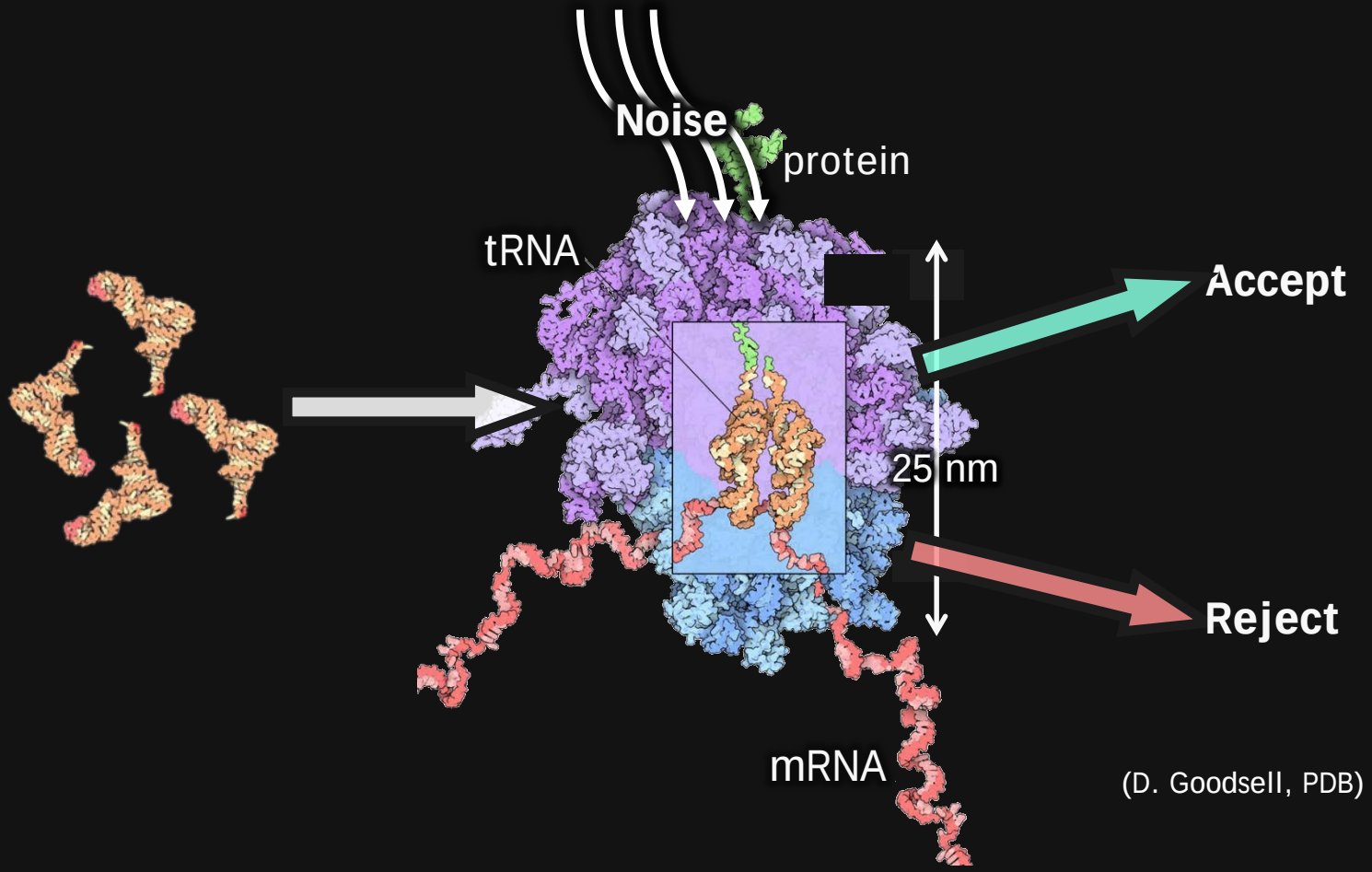
Accuracy effect on fitness?

- Basic protein properties (mutations…)
- Biological context (environment…)





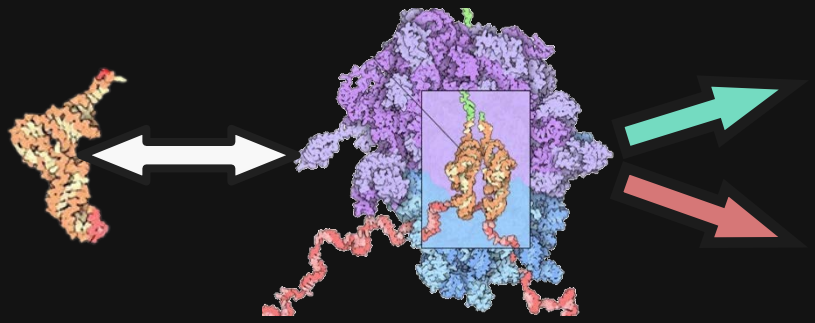
Decoding at the ribosome is a molecular recognition problem

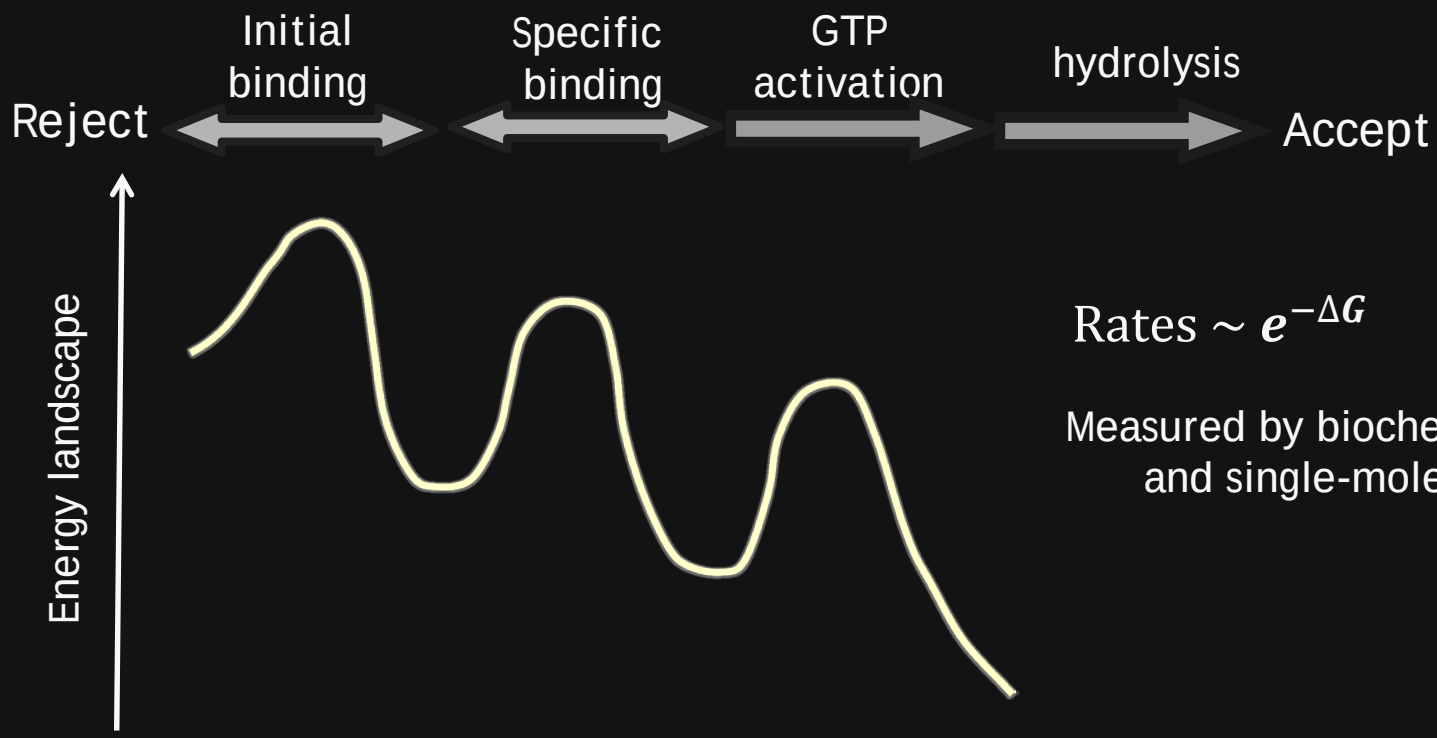


- How to evolve molecules that recognize in a **noisy** environment?
  (crowded, thermally fluctuating, weak interactions).

Decoding governed by energy landscapes of tRNA recognition

- Measured kinetics: sequence of transitions.
- *Large* conformational changes.



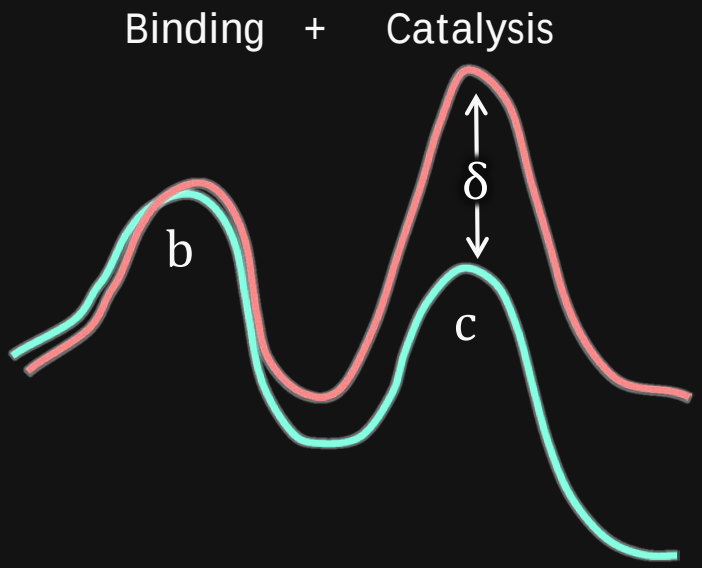


$$\text{Rates} \sim e^{-\Delta G}$$

Measured by biochemistry
and single-molecule experiments.

The ribosome decoding problem is an example of recognition in the presence of competition

- Generic enzyme kinetics (MM):



$$R_B \sim \frac{1}{e^{b} + e^{c+\delta}}$$

$$R_G \sim \frac{1}{e^{b} + e^{c}}$$

- Analysis applies to enzymes: Balance rate R_C and specificity R_G/R_B ?

- What is the recognition performance ("fitness") ?

Recognition “fitness” has generic features

- “Fitness” *F* is obscure and context-dependent.:

  → look for generic properties of $F(R_G, R_B) = F(b, \delta, c)$

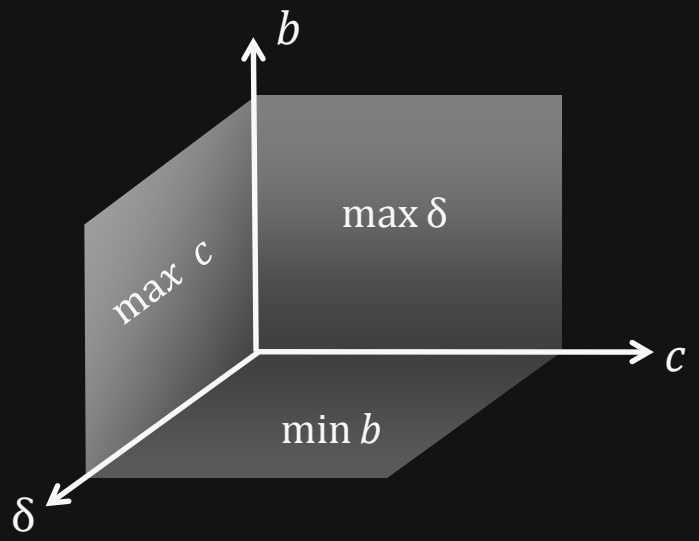
$$R_G \sim \frac{1}{e^b + e^c}$$

$$R_B \sim \frac{1}{e^b + e^{c+\delta}}$$

- “Biologically reasonable”: $\frac{\partial F}{\partial R_G} \geq 0,\ \frac{\partial F}{\partial R_B} \leq 0.$

- Searching for optimum in (b, δ, c) space:

  - **δ approaches biophysical limit.**
  - **Optimization is effectively 1D.**

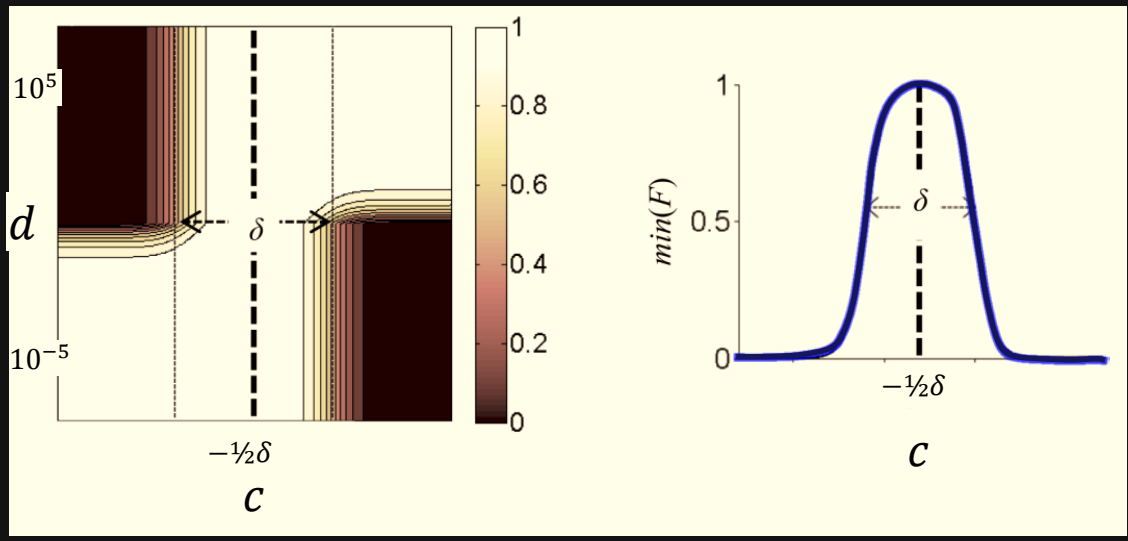


Adaptation in flatland

- Tuning parameter, e.g.

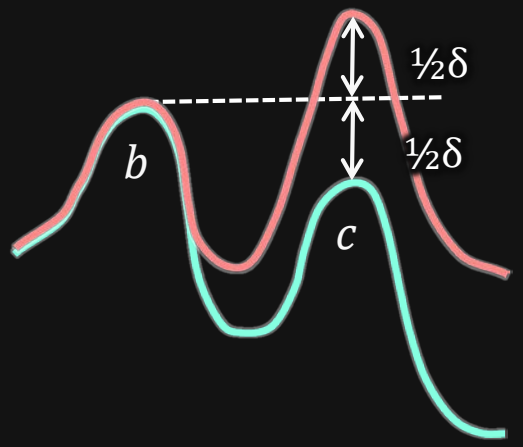
$$F = R_G - d \cdot R_B.$$

"Worst case scenario": $\max(\min(F))$.



- Max-Min solution is **symmetric:**

$$\Delta = c - b \approx -½\delta$$

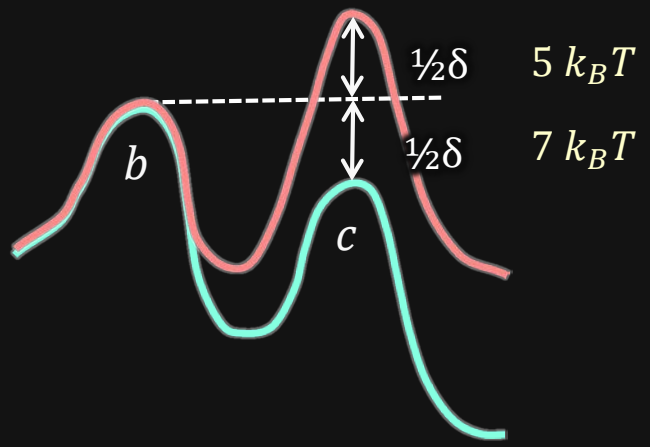


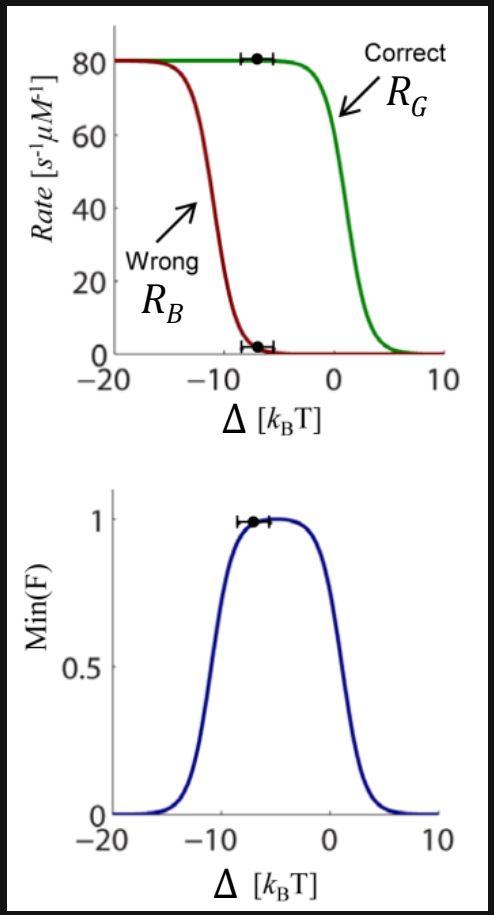
- Any $F(R_G, R_B)$ adapts to Min-Max as long as both rates are "relevant",

$$e^{-\delta} < \left| \frac{\partial F}{\partial R_G} \Big/ \frac{\partial F}{\partial R_B} \right| < e^{\delta}.$$

Ribosome energy barriers appear adapted

- **Max-Min predicts symmetry** $\Delta \approx -\frac{1}{2}\delta$.
- Measurements: $\Delta \approx -7\ k_BT,\ \delta \approx 12 k_BT$.

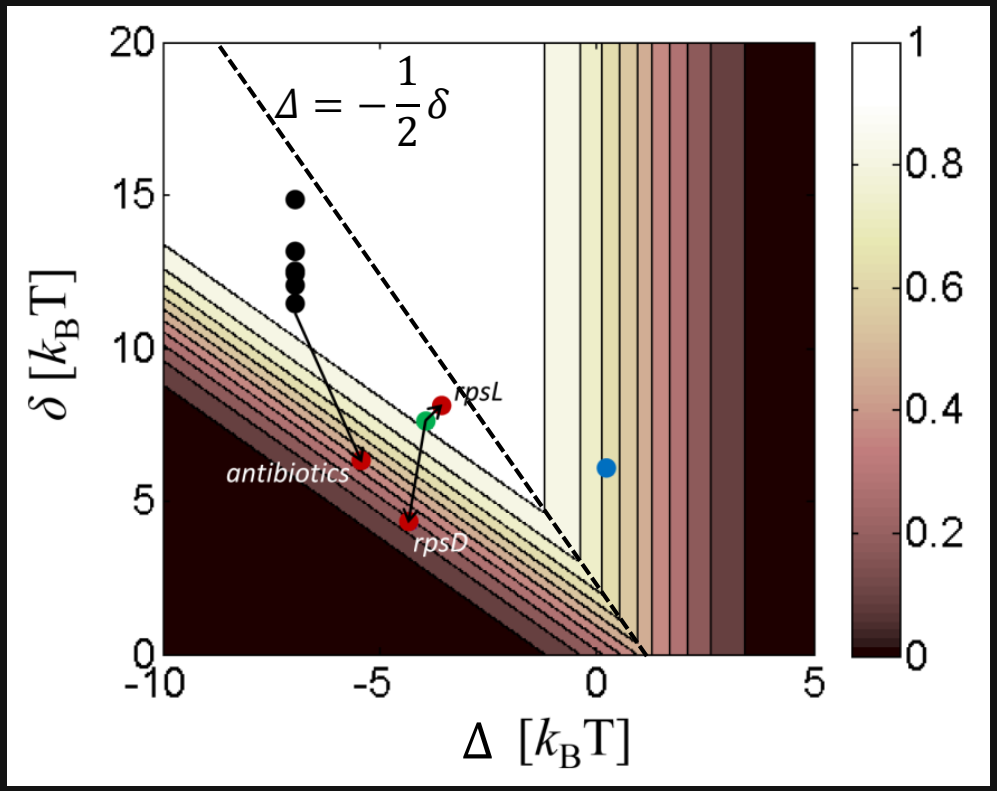


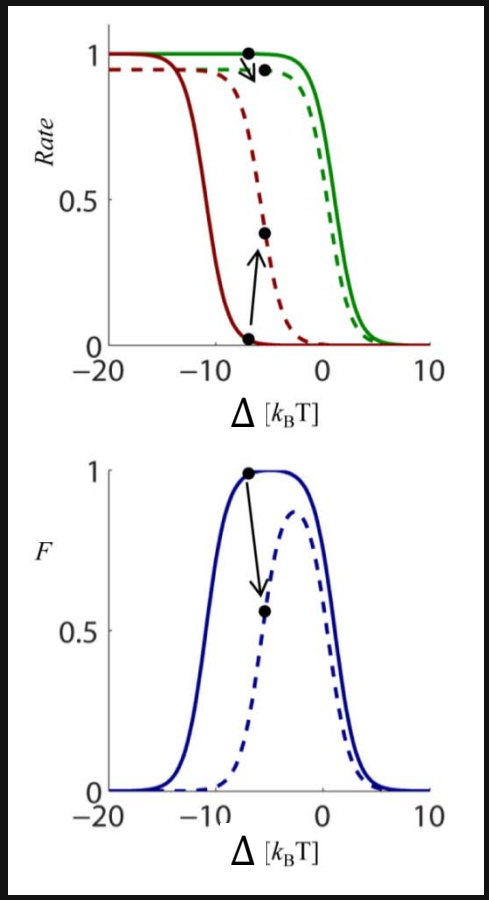


The ribosome shows adaptation

(according to Max-Min prediction).

Predicted adaption regime of other ribosomes





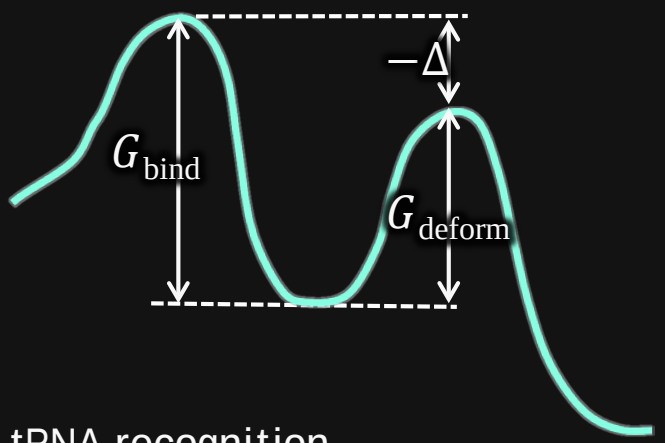
- Adaptation regime in the space of all possible landscapes, $-\delta \leq \Delta \leq 0$.
- Mutations\antibiotics push away from adaptation.

What is the role of conformational changes?

- Energy barrier includes binding and deformation energies:

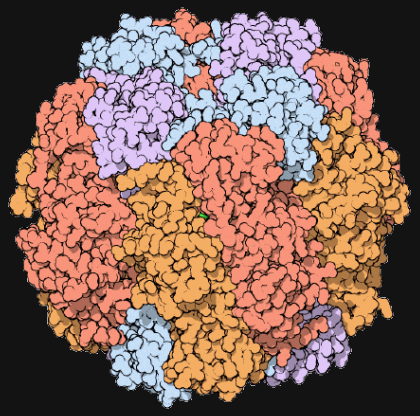
$$\Delta = G_{\text{deform}} - G_{\text{bind}} = -\tfrac{1}{2}\delta \approx 6\ \mathrm{k_B T}.$$

- For any $G_{\text{bind}} \geq 6\ \mathrm{k_B T},\ G_{\text{deform}} \geq 0.$



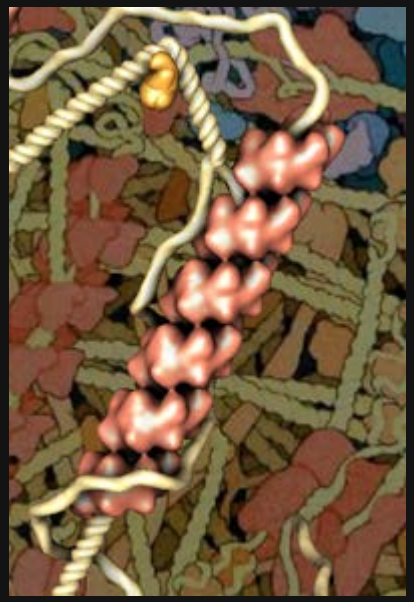
- **Conformational proofreading:** deformation improves tRNA recognition.

Energy barrier that discerns the right target from competitors.



Low-dimensional phenotype space

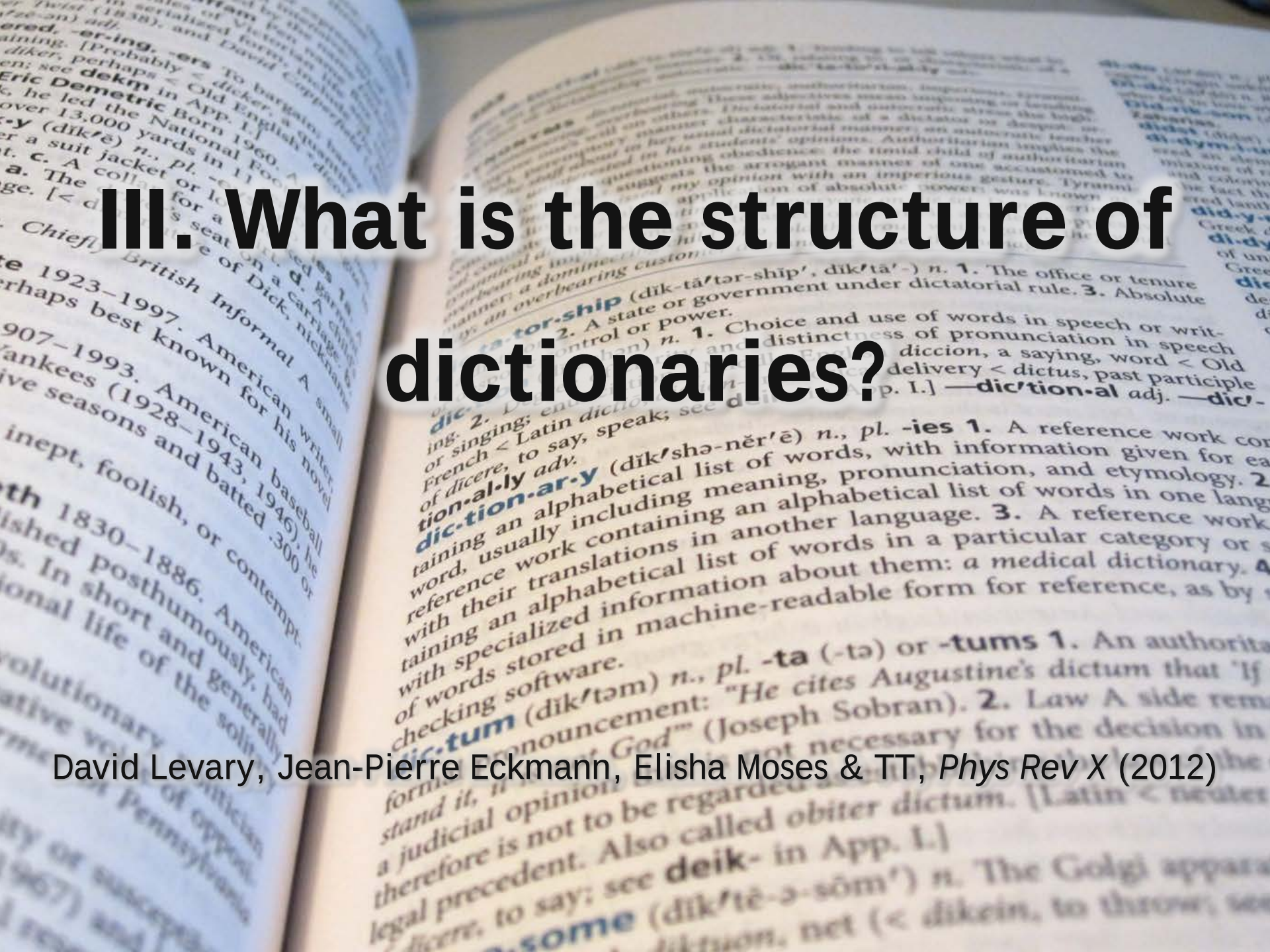
Rubisco proteins adapt to atmosphere (CO_2, O_2) in 1D.



Conformational proofreading

RecA-induced homology search utilizes DNA stretch to detect DNA sequence.

Savir & TT, *PNAS* 2010, *Mol Cell* 2010, *Cell* 2013



III. What is the structure of dictionaries?

David Levary, Jean-Pierre Eckmann, Elisha Moses & TT, *Phys Rev X* (2012)

Dictionaries as Graphs

- New words added to language (dictionaries) to:

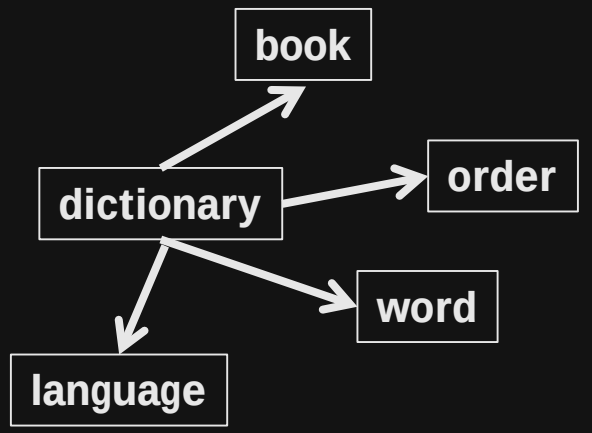
  (i) increase efficiency of communication (ii) introduce new concepts…

- **Dictionary graphs** $G(V, E)$:
  - Vertices V are the defined words.
  - Edges E on $V \times V$ point from word to words in its definition:

    $\{v_1 \rightarrow v_2\} \in E$ iff $v_2 \in D(v_1)$.

Dictionary (n) A *book* which explains or translates, usually in alphabetical *order*, the *words* of a *language*.

D(dictionary) = {book, order, word, language}



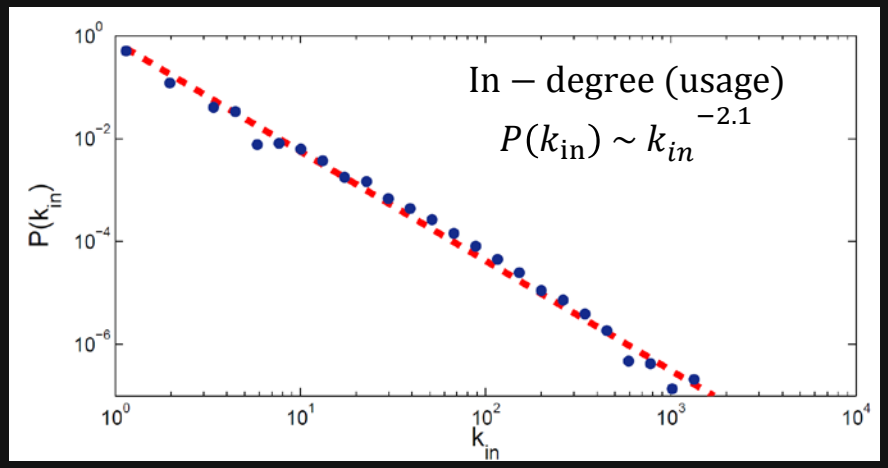
Dictionary graphs are highly connected

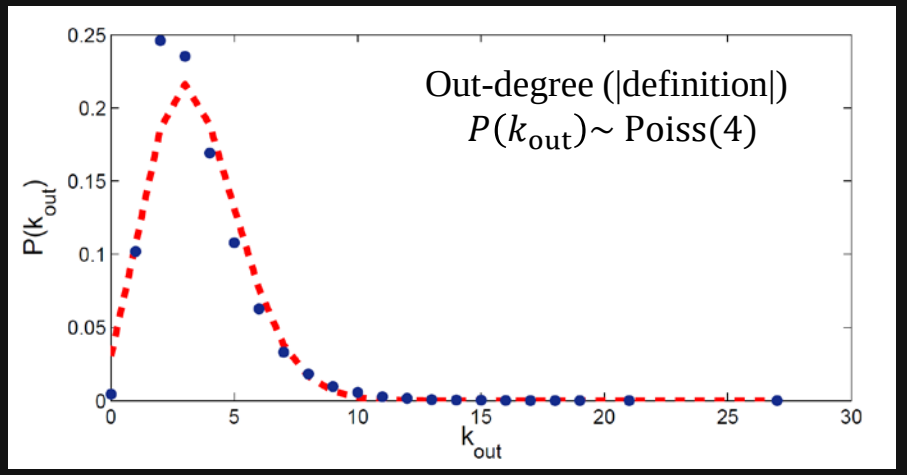
- Main example: (eXtended) **WordNet ®**
- Synsets (grouped synonyms, reduces noise).
- 80,000 synsets (of nouns).
- 300,000 directed edges.
- Results in a large directed graph.

PRINCETON UNIVERSITY

WordNet
A lexical database for English

(Fellbaum et al.)

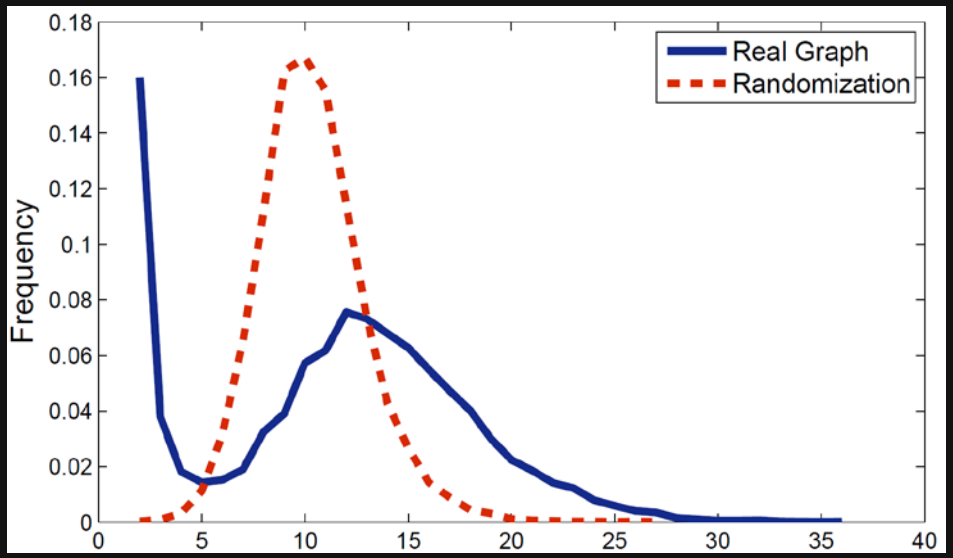




(Dictionary graphs discussed by Litkowski (1970s) and recently by Harnad et al.)

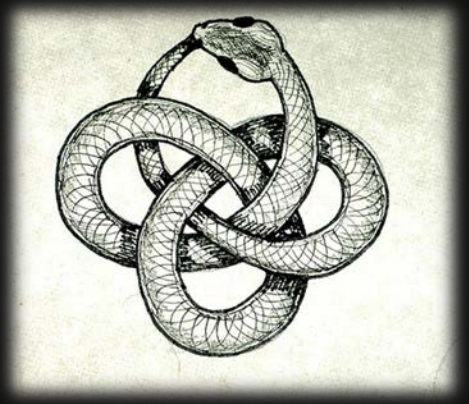
Dictionary graphs contain many short loops

- Compared to randomized graph with same degrees.



?

Theoretical motivation: why dictionaries are loopy?



- **Loops**: from classical paradoxes to self-replication in biology…
  related to emergence…
- **Dictionaries** are inherently cyclic (all words are defined),
  much more than randomized graphs.

- **Possible roles of loops in structure and evolution of dictionaries (languages)?**

- **Suggestion: Loops are not artifacts but related to emergence of *new concepts*.**

- **Philosophical problem: What is a concept?**

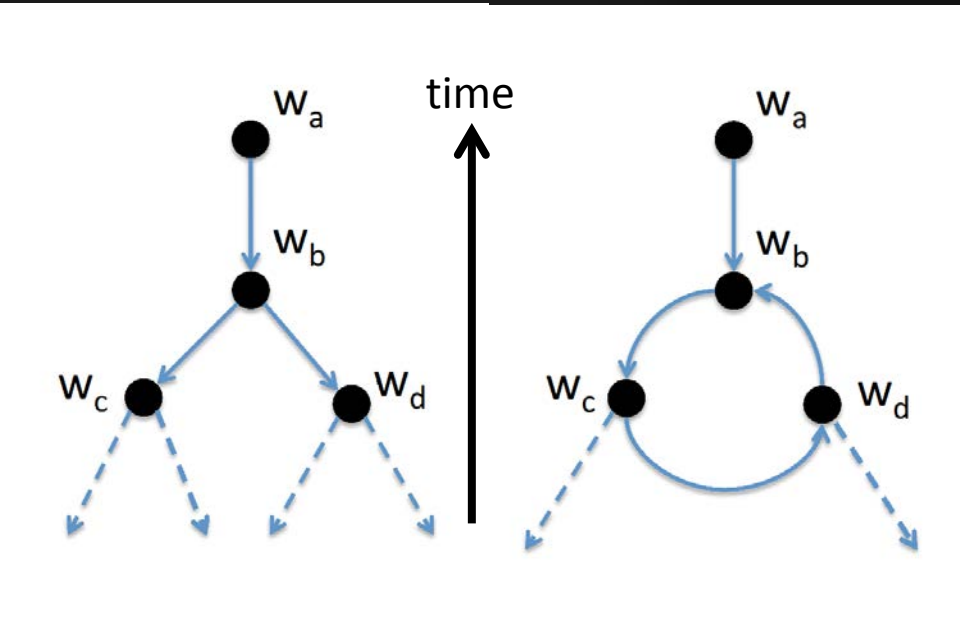
Is emergence of loops also emergence of concepts?

- Emergence of new "concept" at t_0 ≡ word *w* was *undefinable* at $t < t_0$.

- *D*(w) = {words in definition of *w*}.
- *D*(*D*(*D*(*D*...)))) = walking along definitions.

- Word can be replaced by definition:

$$w \to D(w) \to \cdots \to D^{(n)}(w).$$

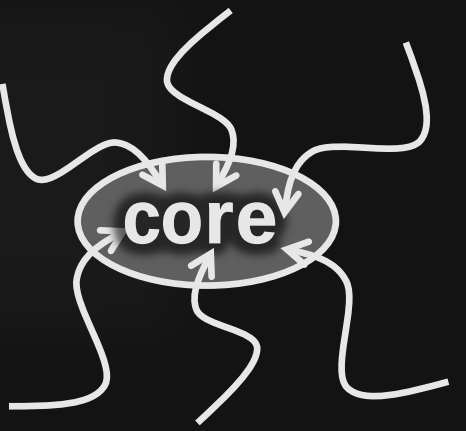


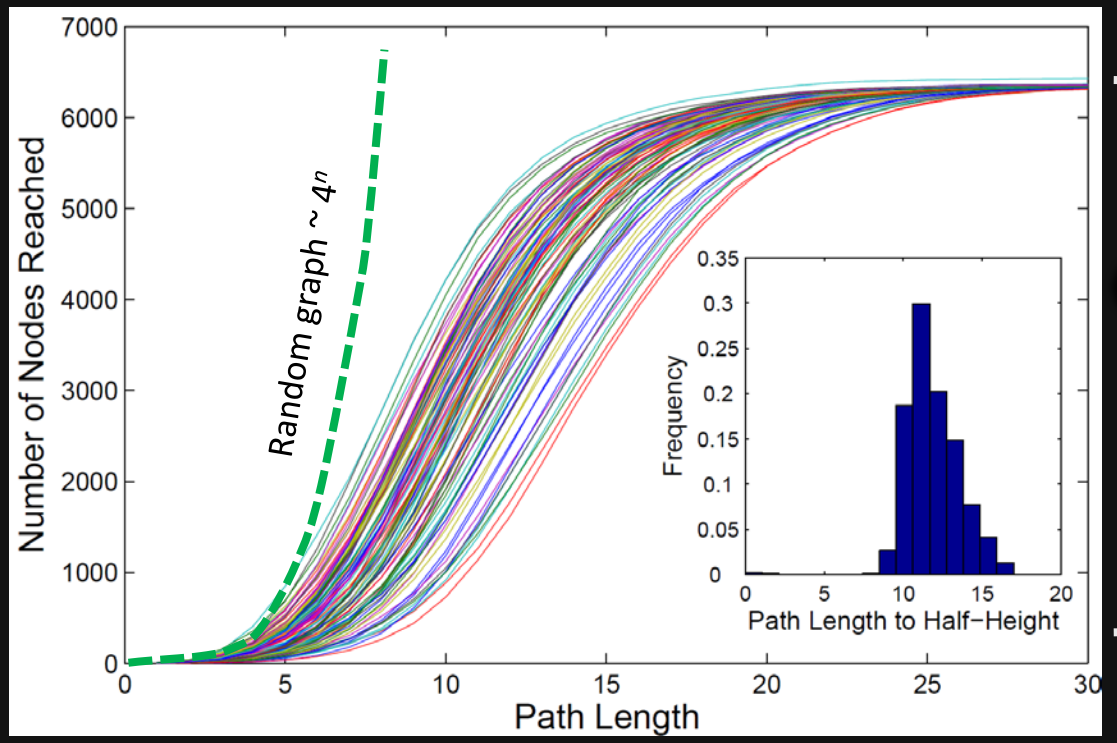
- *G* is tree: translate to older words, everything is definable.
- Loop cannot be reduced to older words: undefinable words → new concepts.

The dictionary converges into a "core"

- eXtended WordNet resolves polysemous words (different meanings).
- Examine only nouns (most related to concepts).
- Definition paths converge *fast* into "**core**" =

  ~ 6,300 words reachable from 99% of words.
- Graph is (i) *not tree-like* (ii) *not small-world* (large diameter).





Examples of strongly connected components

emotion	height	bark	injury	winner
spirit	end	trunk	violence	contestant
dejection	dimension	tree	accident	competition
melancholy	length	lumber		
feeling				

Cores are quite universal and overlap with most useful words

	Core	Ogden (600)	Jōyō Kanji	Gutenberg
Core	1595	314 (52%)	403 (29%)	265 (39%)
Basic English		600	328 (24%)	213 (32%)
Jōyō Kanji			1376	319 (47%)
Gutenberg				673

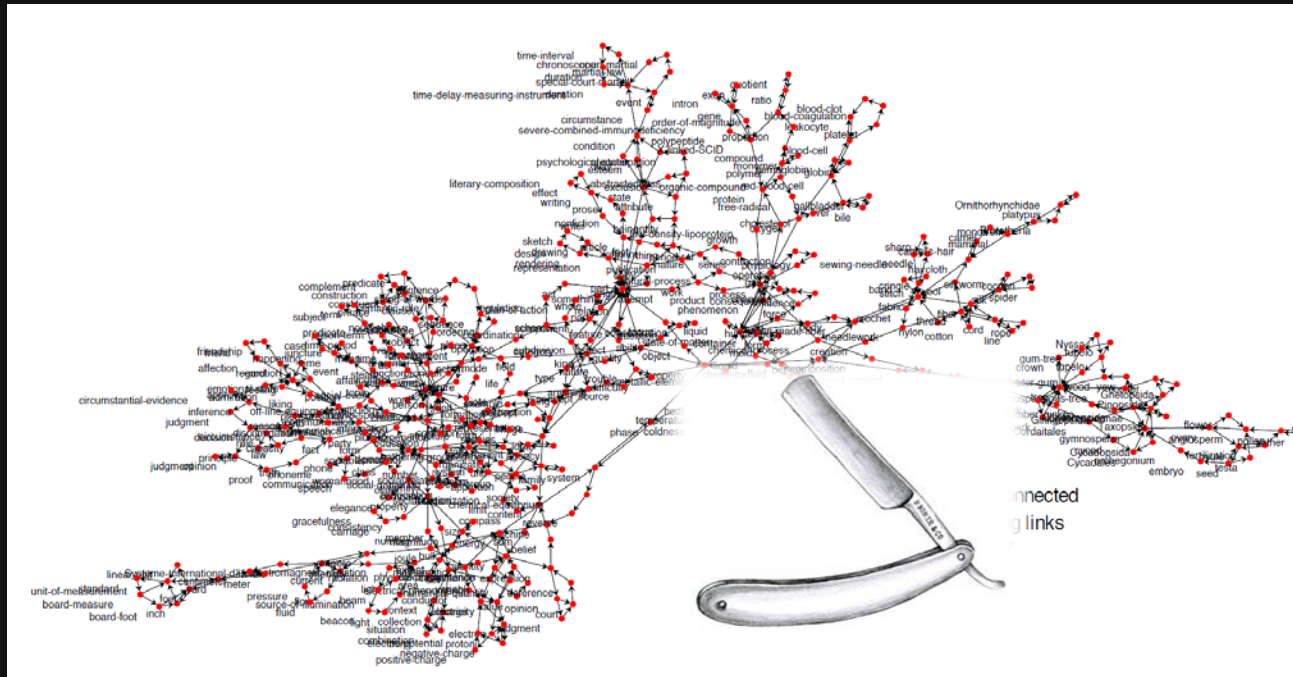
- Ogden's **"Basic English":** 850 words for daily discourse (600 nouns)
- **Joyo Kanji**: Japanese Education Ministry's list of essential 1,945 characters (1376 nouns).
- **Gutenberg**: top 1000 words in all books on Project Gutenberg (673 nouns)

- Missing words are "over-specific" yet useful words: apple, brick, chalk…

$$\text{Core} \not\ni \text{apple} \in \text{Ogden}, \quad \text{but } D(\text{apple}) = \{\text{red}, \text{fruit}\} \in \text{Core}.$$

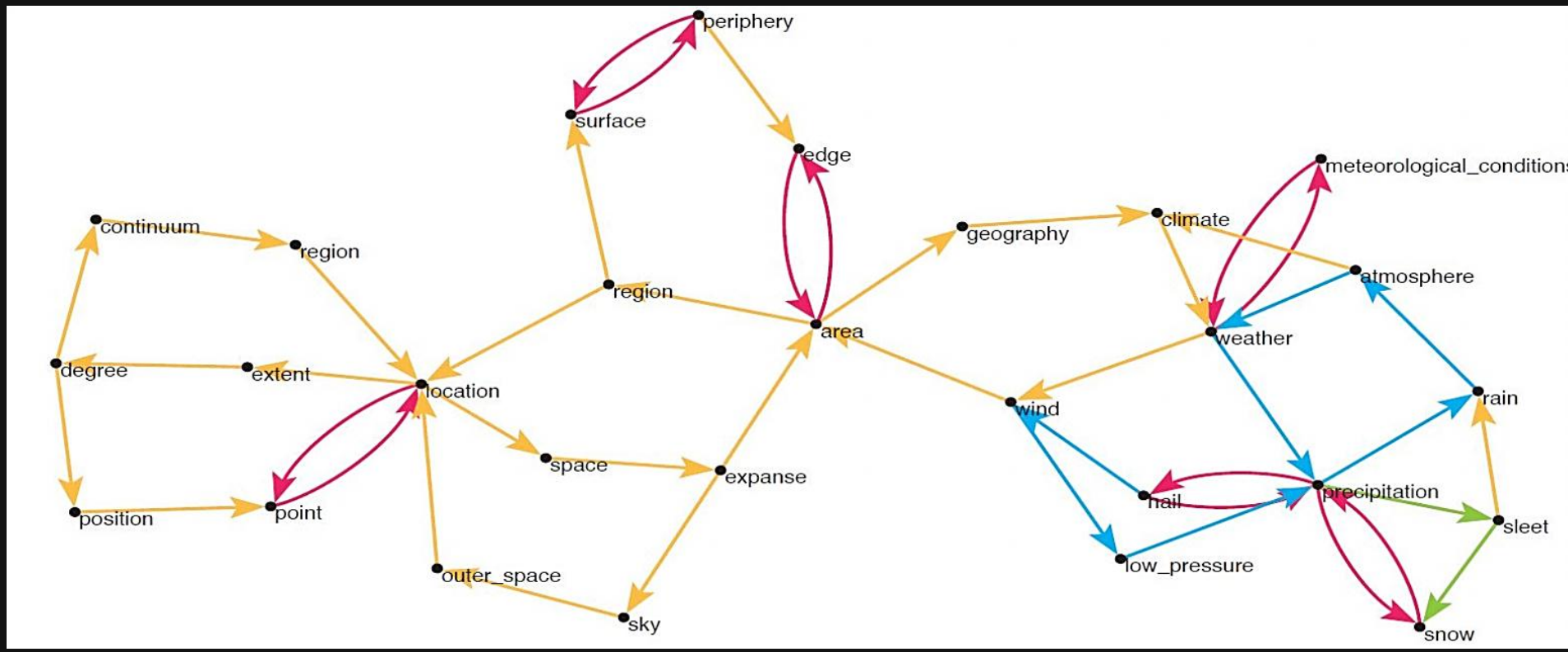
The core is made of loops

- **Core** is obtained by "*shaving hairs*" (words not used in definitions).
- Shaving seven times leaves only loops.
- ~9000 words are in loops.
- ~1,100 *strongly connected component* (everybody reaches everybody).
- Most connected components are small but one has 6300 nodes = the "core".



The loop structure of the core

- Example of a large strongly connected component (SCC):



- Red=2, Green=3, Blue=4, Yellow=5.

Concept “basis”? Large-scale structure of concepts?

- Possible “decomposition” of words:
  - Cut links on large loops (>5): core decomposes ~400 clusters.
  - Proximity of each word to each cluster is measured by # of short paths.
- Proximity matrix 80000 x 400 → SVD → “eigenmodes” of loop clusters.
- eigenmodes are semantically cohesive clusters of words from decomposed core.

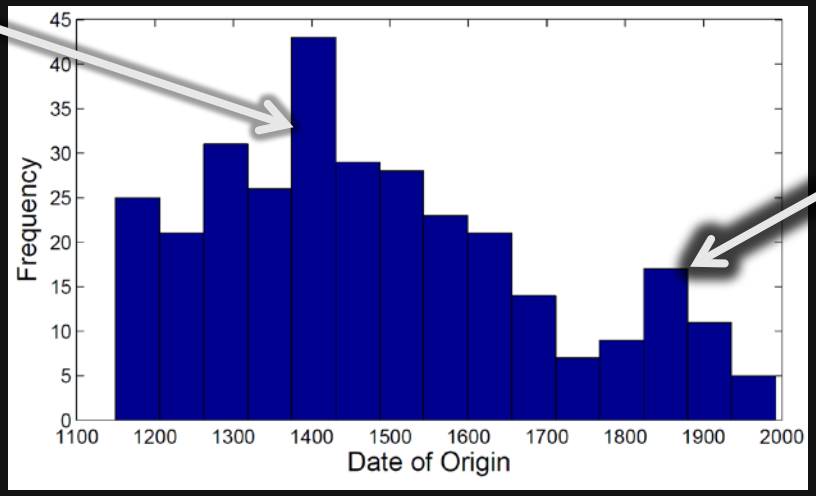
Vector 1	Vector 2	Vector 5	Vector 6	Vector 8
Old World, oceans	*spine, brain*	grains	*body parts*	Jesus, Christianity
bodies of water	*body parts*	*bodies of water*	*flower, seed*	man, woman
The Americas	Old World, oceans	herbage	*tree, bark*	*nucleus, DNA*
influence, power	*energy*	flower, seed	grains	Roman Empire
	pathology	*land*	nucleus, DNA	student, teacher
	narrative	*water*	spine, brain	speaker, speech
	cognition		*Gymnosperms*	Vatican, absolution
	organic process		bodies of water	Old Testament
	respiration		pathology	book

Dynamics of dictionary construction

- Date of word's first appearance (online etymology dictionary).
- Approximated order of construction $W = \cup_t W_t$.

- Temporal loop distribution exhibits two distinct peaks:

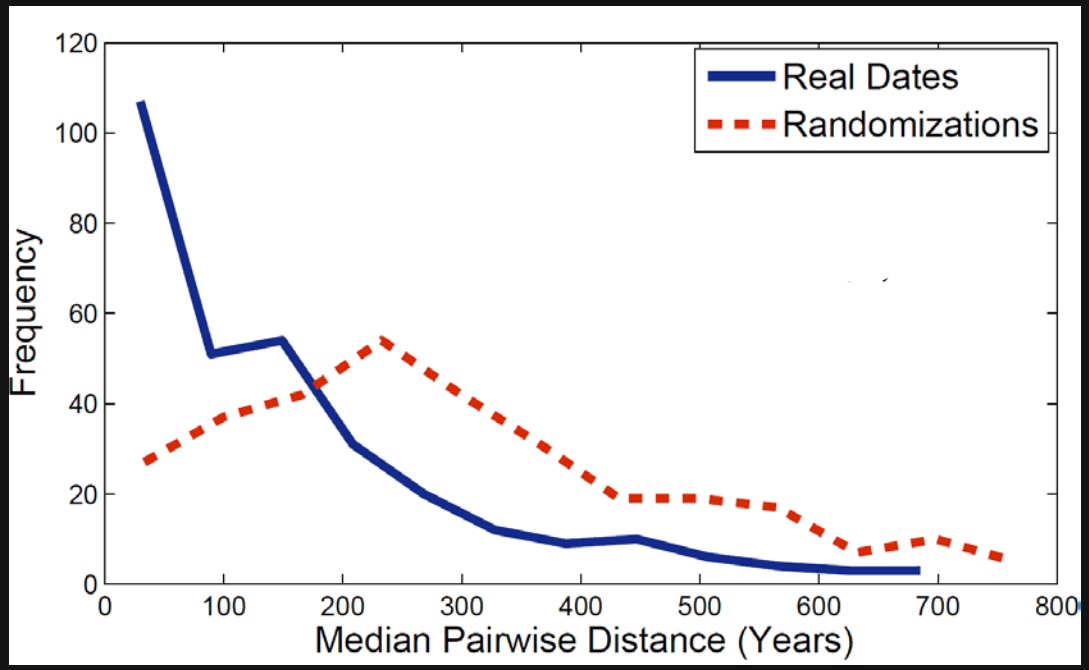
14th-16th century:
Middle → Modern English



19th-20th century:
science & technology

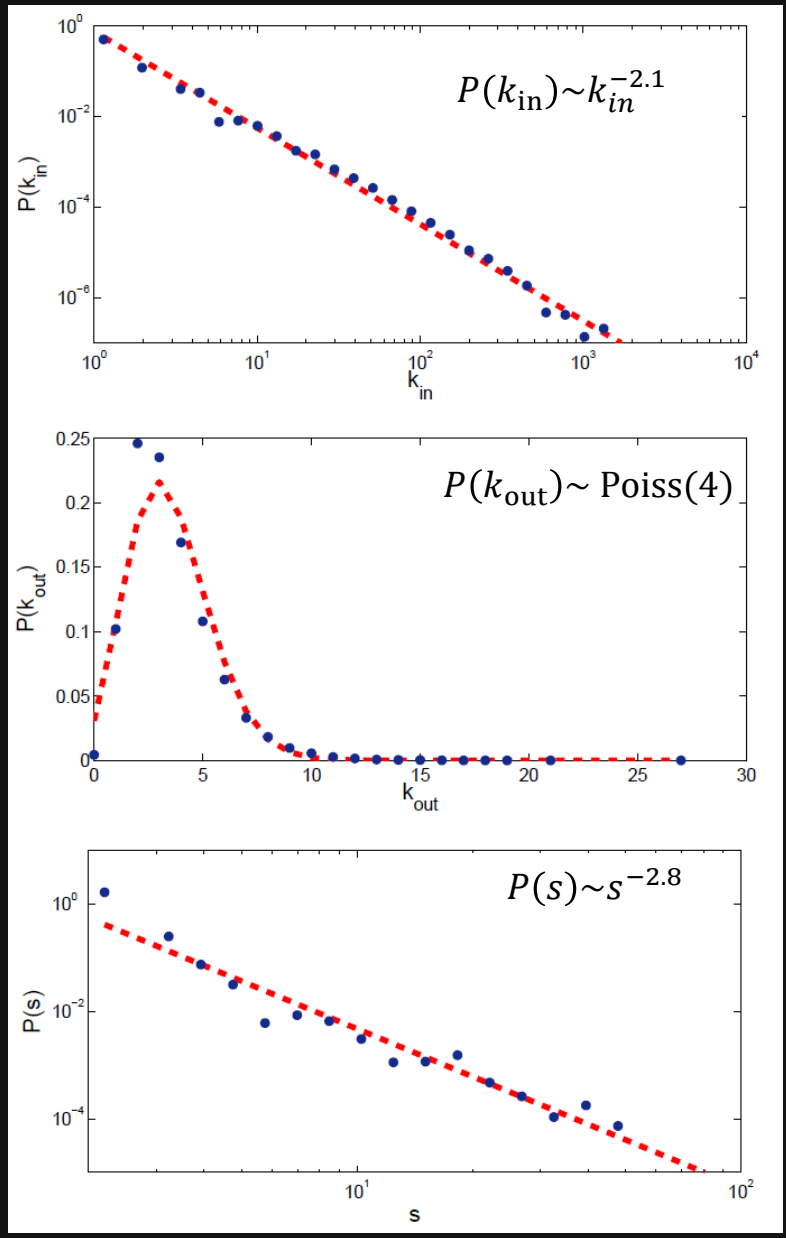
Loops tend to appear “simultaneously”

- Median pairwise distance between words in a loop.



Scale-free loop distribution

- Distribution of strongly connected component size: $P(s) \sim s^{-2.8}$.



Toy model "explains" Scale-free loop distribution

- Preferential attachment model: definitions tend to use more common words.

- Toy model: connect new edges from a new words in proportion to $a + k_{in}$

  (a > 0 is the "attractiveness" of newly introduced word).

$$P(k_{\text{in}}) \sim k_{\text{in}}^{-(2+\delta)}$$

$$\delta = \frac{a}{\langle k_{\text{out}} \rangle}$$

- Adding loops:
- Loops start new SCC (p) or join existing SCC (1-p).

$$P(s) \sim s^{-[1+1/(1-p)]}$$

- Consistent with $\gamma = 2.8$ for $p = 0.4$

$$P(s) \sim s^{-[1+1/(1-p)]}$$

Summary

- Short loops ~ simple semantic units of dictionary graphs.
- Highly connected “core” of basic words.
- Overall structure of loops reflects semantic features, esp. if loops are labeled by another “coordinate” like "time“ (dynamics) which adds structure to the topology of dictionaries.
- Suggests relation to meaning: New loops ↔ New “concepts”.
- Tend to appear “simultaneously” (following historical dictionaries)
- SVD: eigenmodes of loops (concepts)?

Thanks

Molecular information systems

Yonatan Savir (Harvard)

Elad Noor & Ron Milo (Weizmann)

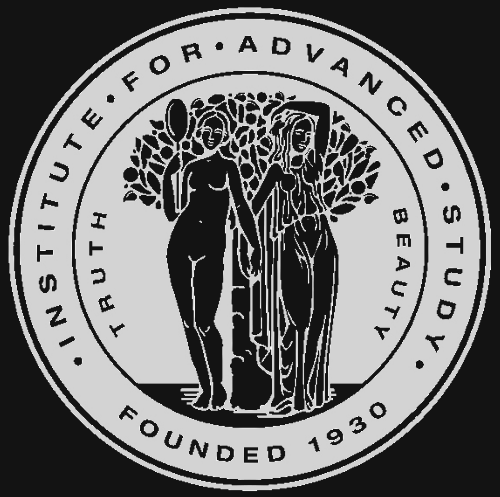
Dictionaries

Jean-Pierre Eckmann (Geneva)

David Levary (Harvard)

Elisha Moses (Weizmann)

Christiane Fellbaum (Princeton)



Directions

- Other languages, multi-lingual mapping…
- Association network (also very loopy).
- Spectral properties…
- Components of higher dimension (simplexes?)…
- Biochemical networks (mostly core).