Generalized Topological Spaces in Evolutionary Theory and Combinatorial Chemistry

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science & Interdisciplinary Center for Bioinformatics, **University of Leipzig**
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

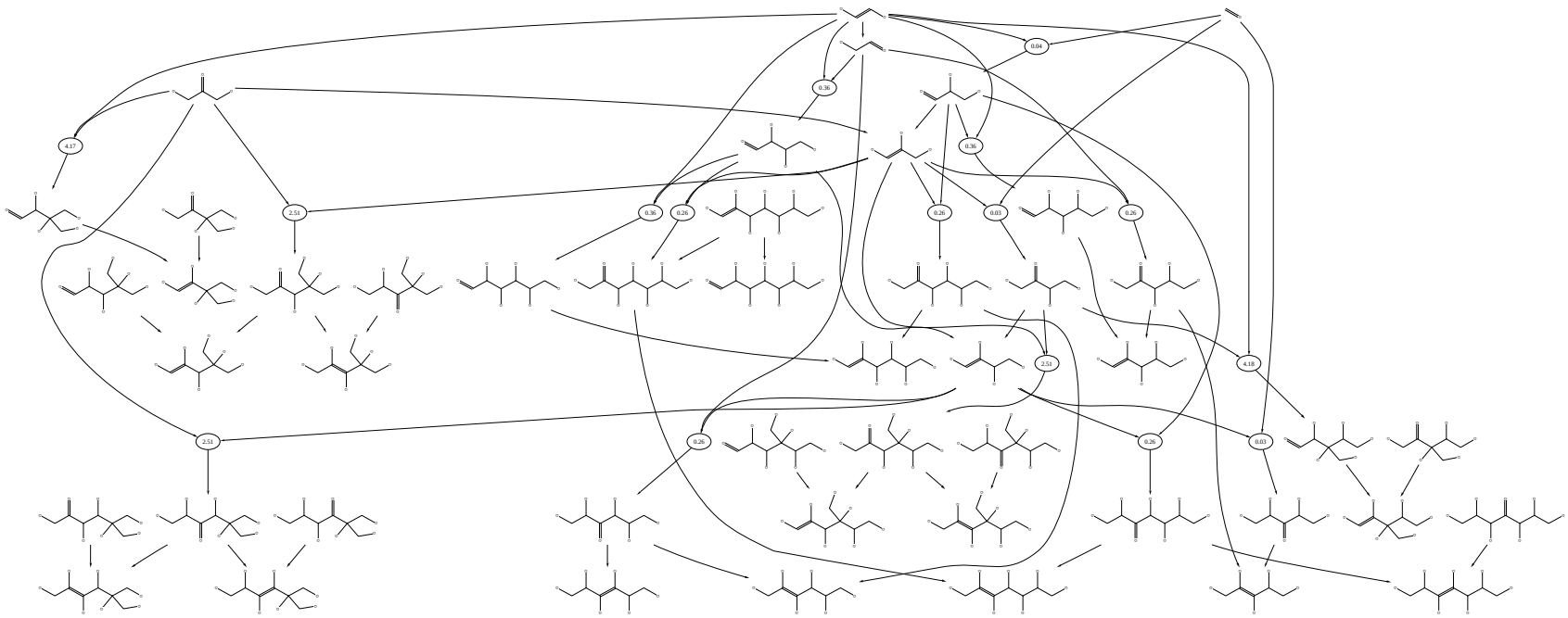
Joint work with
Bärbel M. R. Stadler
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Bled, Feb 20 2006

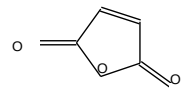
Outline

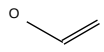
- The Space of Molecules
- Topological Theory in Molecular Evolution
  - Growing Populations
  - Abstract Closure Spaces
  - Separation, Continuity, and Products
  - A Theory of Characters (Merkmale)
- Fitness Landscapes
  - Local Minima
  - Connectedness
  - Barrier Trees
- Chemical Organizations
  - Chemical Reactions and Stationary Fluxes
  - Dittrich's Theory
  - Chemical Organization as Topology

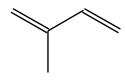
Chemical Reaction Networks

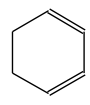


A Growing Chemical Network



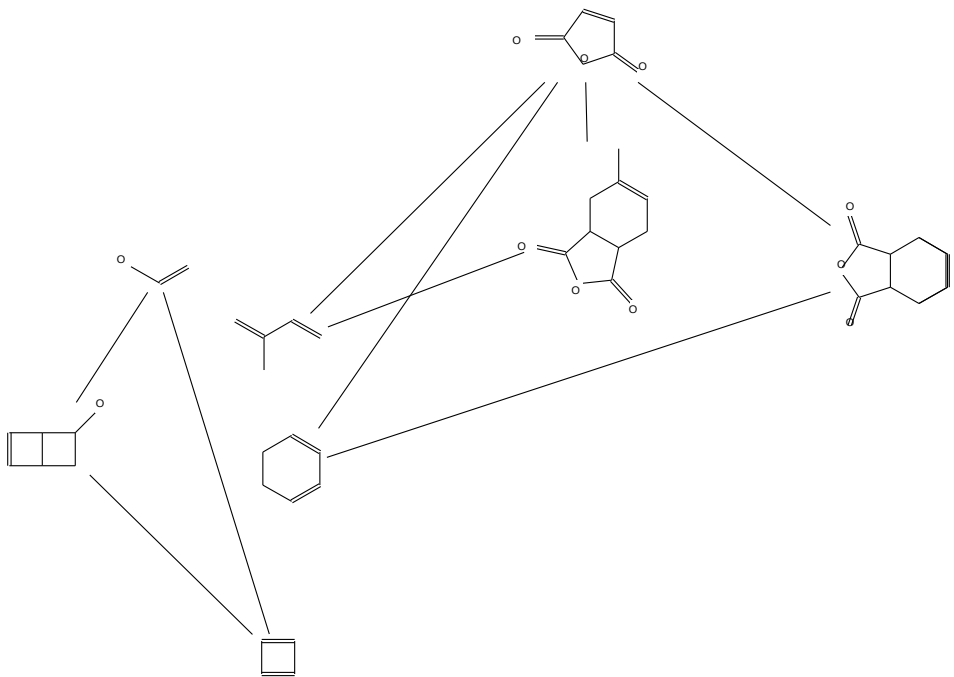






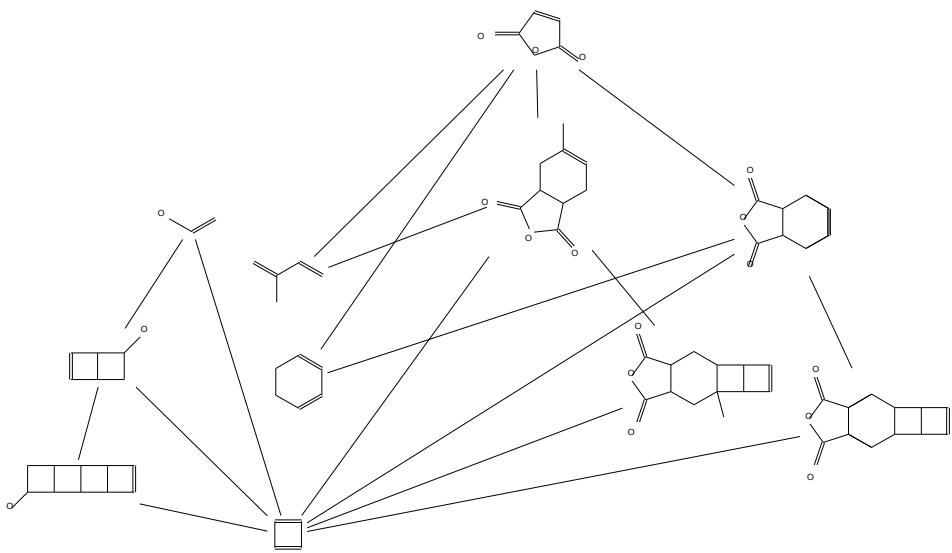
A

A Growing Chemical Network



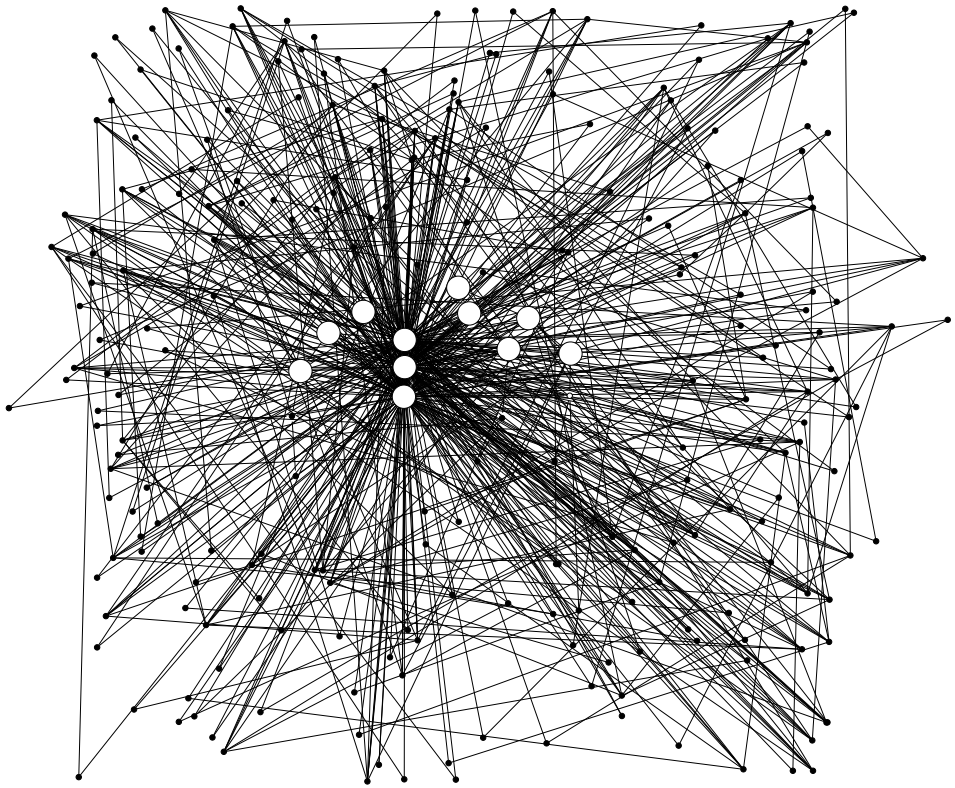
$\mathrm{cl}(A)$

A Growing Chemical Network



$\mathrm{cl}(\mathrm{cl}(A))$

A Growing Chemical Network



$\mathrm{cl}(\mathrm{cl}(\mathrm{cl}(A)))$

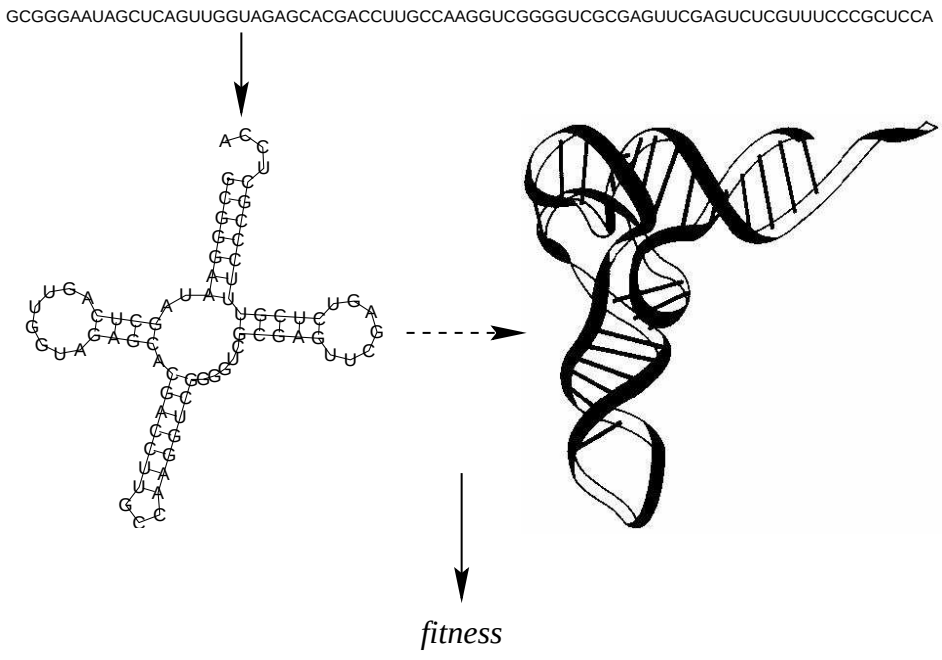
“Chemical Space”

* Chemical Reactions define a mathematical structure of “accessibility” on the set of molecules

?? What is the (mathematical) structure of the operator cl?

Before we get to this . . .
. . . let’s have a look at a very different topic!

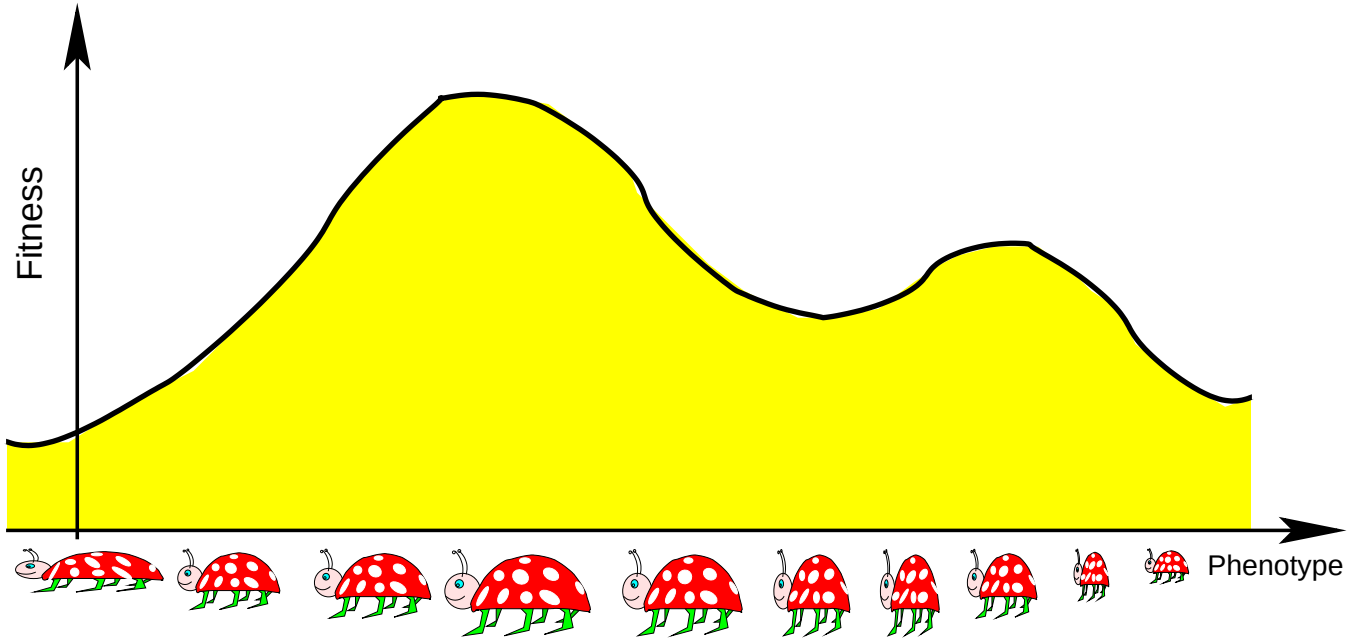
Evolutionary Biology: Genotype-Phenotype Maps



Given:

- A **set** X of genotypes (sequences)
- Genetic Operators (mutation and/or recombination ...)
- A **set** Y of (potential) phenotypes (structures)
- A **function** $f : X \longrightarrow Y$ assigning a phenotype to each genotype.

The "usual" view

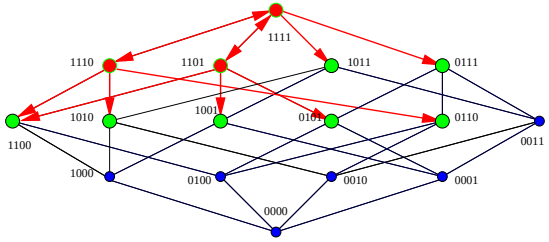


- Phenotypes are "somehow" numbers or *vectors*
- Accessible phenotypes are within a small (Euclidean) distance (in this vector space)
- Fitness is a (more or less) smooth function

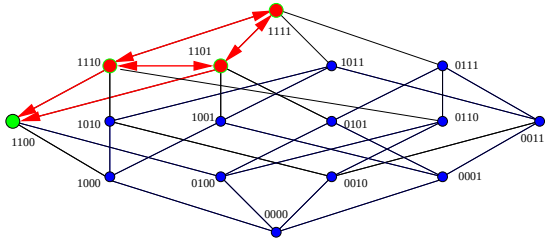
Population Genetics is perfectly happy . . .

So, what is wrong with this picture?

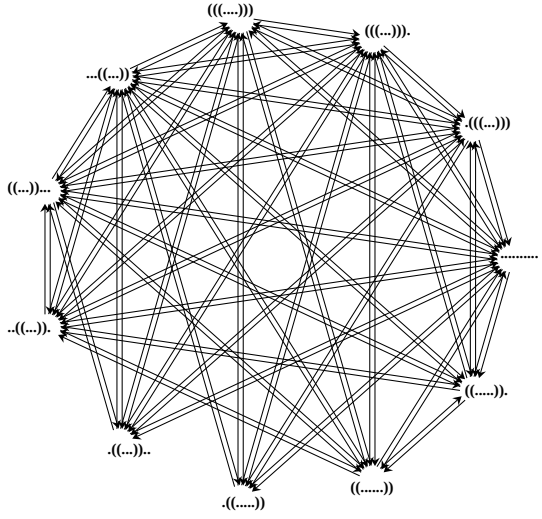
(1) Genotype space is **discrete**



(2) Its structure depends on the **genetic operator**



(3) Phenotype space inherits its structure from genotype space



Accessibility at genotypic levels **implies** accessibility at phenotypic level

Does Evolution really "live" on an Euclidean space????

Goal: A "Relative" Theory

We want a theory of phenotypes that can deal with concepts such as

- Continuity and Discontinuity
- Character
- Homology
- Innovation

WITHOUT recourse to a

specific representation of the phenotype

Genotype Spaces

Given:

a set X of possible genotypes

a set A of realized genotypes

a fixed collection of genetic operators

[such as mutation, recombination, gene-rearrangement]

define the set A' of genotypes accessible from A.

Properties

(i) No spontaneous creation, i.e, $\emptyset' = \emptyset$.

(ii) A more diverse population produces more diverse offsprings: $A \subseteq B$ implies $A' \subseteq B'$

(iii) All parental genotypes are also accessible in the next time step $A \subseteq A'$.

This is the same as the C-operator for the chemical network!

In the case of mutation as the only source of diversity:

haploid populations, no sex, no recombination, etc

(iv) Diversity of offsprings depends only on the parent: $A' = \bigcup_{x \in A} \{x\}'$

Generalized Closure Spaces

... instead of vector spaces ...
Set X, closure function $\mathrm{cl} : \mathcal{P}(X) \to \mathcal{P}(X)$
Equivalent formulations:

$$\mathrm{int}(A) = X \setminus \mathrm{cl}(X \setminus A)$$

The interior is the dual of the closure function
A set N is a neigbhorhood of x if and only if $x \in \mathrm{int}(N)$.
Let $\mathcal{N}(x)$ be the set of all neighborhood of x.
$\mathcal{N} : X \to \mathcal{P}(\mathcal{P}(X))$.
Closure, interior and neighborhood functions are equivalent.

Generalized Closure Spaces

	closure	neighborhood
K0	$\mathrm{cl}(\emptyset) = \emptyset$	$X \in \mathcal{N}(x)$
K1	$A \subseteq B \implies \mathrm{cl}(A) \subseteq \mathrm{cl}(B)$ $\mathrm{cl}(A \cap B) \subseteq \mathrm{cl}(A) \cap \mathrm{cl}(B)$ $\mathrm{cl}(A) \cup \mathrm{cl}(B) \subseteq \mathrm{cl}(A \cup B)$	$N \in \mathcal{N}(x),\ N \subseteq N'$ $\implies$ $N' \in \mathcal{N}(x)$
K2	$A \subseteq \mathrm{cl}(A)$	$N \in \mathcal{N}(x) \Rightarrow x \in N$
K3	$\mathrm{cl}(A \cup B) \subseteq \mathrm{cl}(A) \cup \mathrm{cl}(B)$	$N', N'' \in \mathcal{N}(x) \implies$ $N' \cap N'' \in \mathcal{N}(x)$
K4	$\mathrm{cl}(\mathrm{cl}(A)) = \mathrm{cl}(A)$	$N \in \mathcal{N}(x) \iff$ $\mathrm{int}(N) \in \mathcal{N}(x)$
K5	$\bigcup_{i \in I} \mathrm{cl}(A_i) = \mathrm{cl}\left(\bigcup_{i \in I} A_i\right)$	$\mathcal{N}(x) = \emptyset$ or $\exists N(x) : N(x) \subseteq N$ iff $N \in \mathcal{N}(x)$

In general: only (K0), (K1), (K2) hold: **neighborhood space** For mutation in haploid populations: (K0), (K1), (K2), (K5) [and thus (K3)]: **additive pretopological space**

For comparison: (K0), (K1), (K2), (K3), and (K4) are equivalent to the axioms of a **topology**.

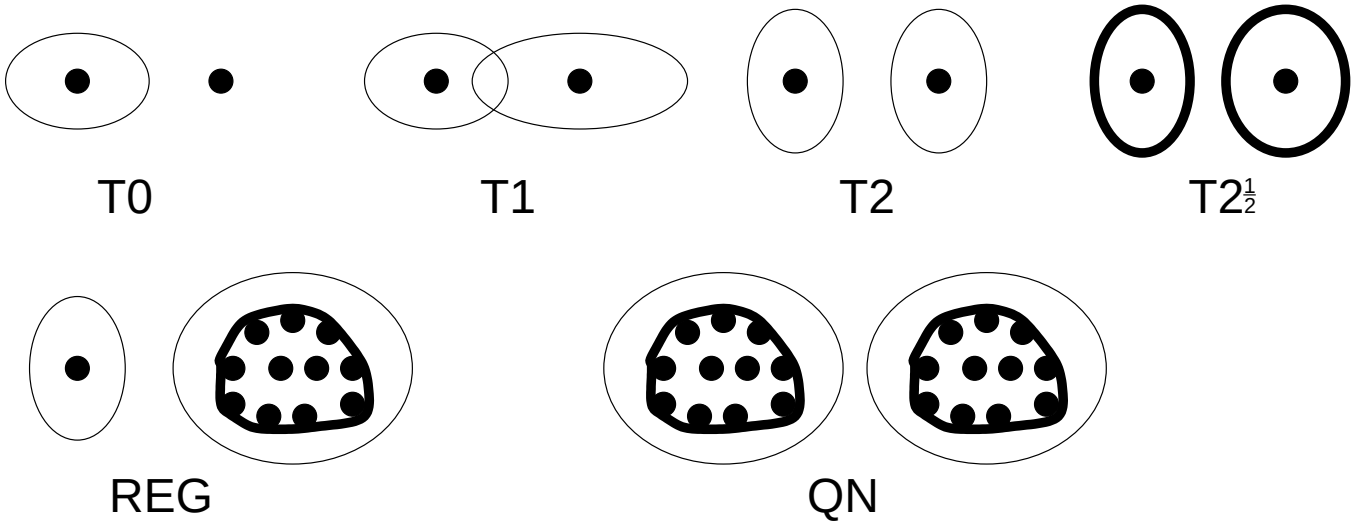
Cool!

. . . so, real evolution, genetic algorithms, evolution strategies, multi-objective optimization heuristics, genetic programming, etc., etc., live on a neighborhood space.
. . . for mutation only, it is even a pretopology.
Thus:
Directed graphs and finite pretopological spaces are the same thing

Should we care that our closure function is NOT idempotent?
NO, Eduard Čech in the 1960s wrote a big, fat textbook on point set topology, where he showed that pretty much everything works in pretopologies — thus you can do topology without every talking about open or closed sets. (Just the proofs get a bit more tedious without this convenience.)

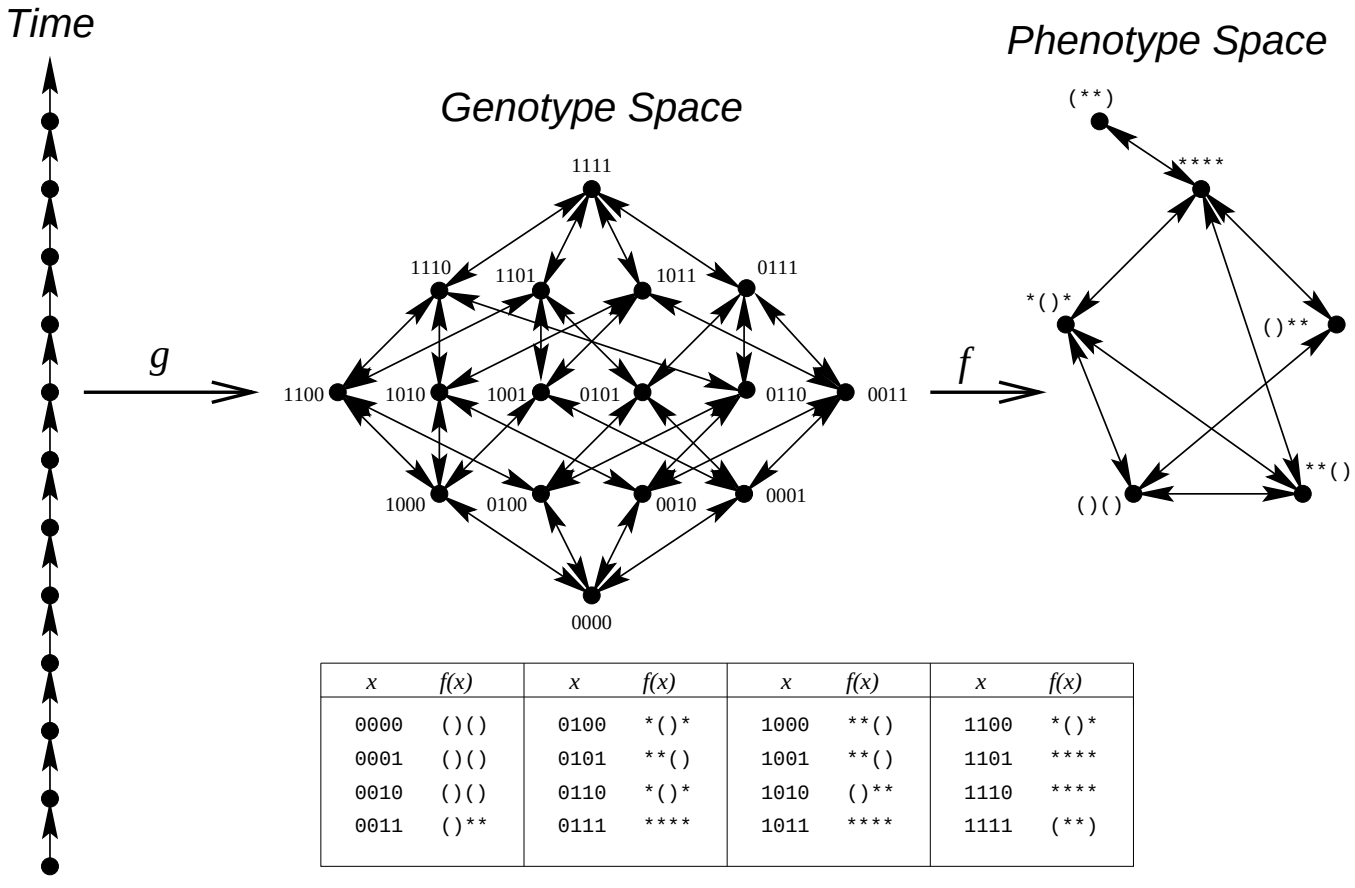
What can we prove about closure spaces?

For example: Implications between *Separation Axioms*



$$(\text{T4}{=}\text{T0}{+}\text{R0}{+}\text{QN}) \implies (\text{T3}{=}\text{T0}{+}\text{REG}) \implies \text{T2}\tfrac{1}{2} \implies \text{T2}$$
$$\implies \text{T1} \implies \text{T0}$$

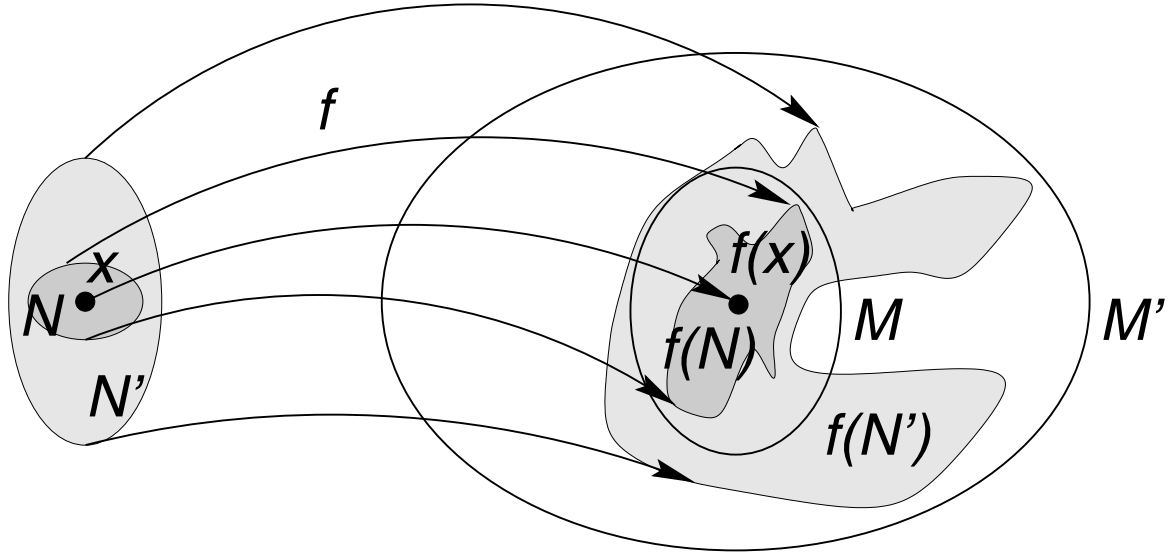
Evolutionary Trajectories



x	f(x)	x	f(x)	x	f(x)	x	f(x)
0000	()()	0100	*()*	1000	**()	1100	*()*
0001	()()	0101	**()	1001	**()	1101	****
0010	()()	0110	*()*	1010	()**	1110	****
0011	()**	0111	****	1011	****	1111	(**)

Continuity

Genotype-Phenotype map: $(X, \mathrm{cl}) \to (Y, \mathrm{cl})$



Equivalent: **closure preservation**: $f(\mathrm{cl}(A)) \subseteq \mathrm{cl}(f(A))$.

BUT: What is closure in phenotype space?
$y' \in \mathrm{cl}(B)$ ... y' is "readily accessible" from B
i.e., there are "enough" genotypes that fold into members of B who can mutate or recombine into an offspring with phenotype y'.
$\Rightarrow$ Notion of continuous versus discontinuous trajectories

Product Spaces

Let (X_1, c_1) and (X_2, c_2) be two general closure spaces, with neighborhood systems $\mathcal{N}_1$ on X_1 and $\mathcal{N}_2$ on X_2.

Product space: Point set $X_1 \times X_2$.

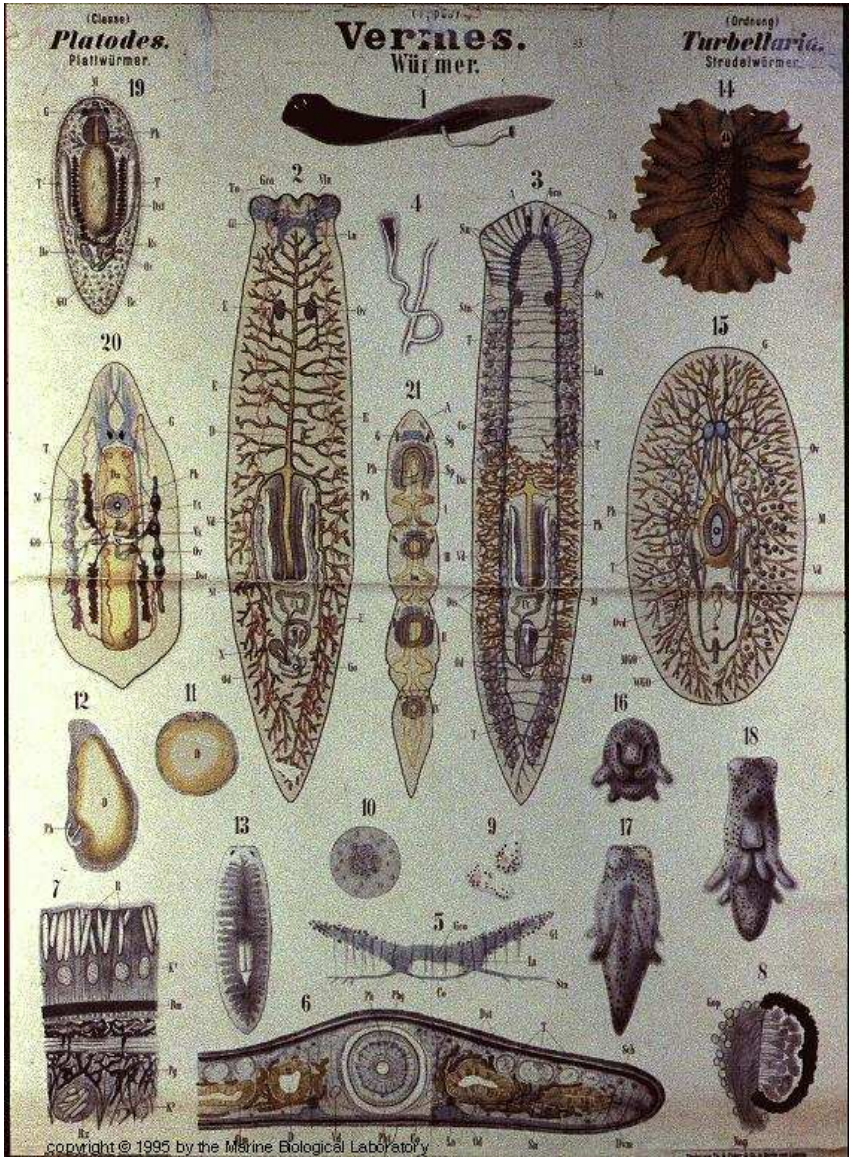
The neighborhoods of the product space satisfy:

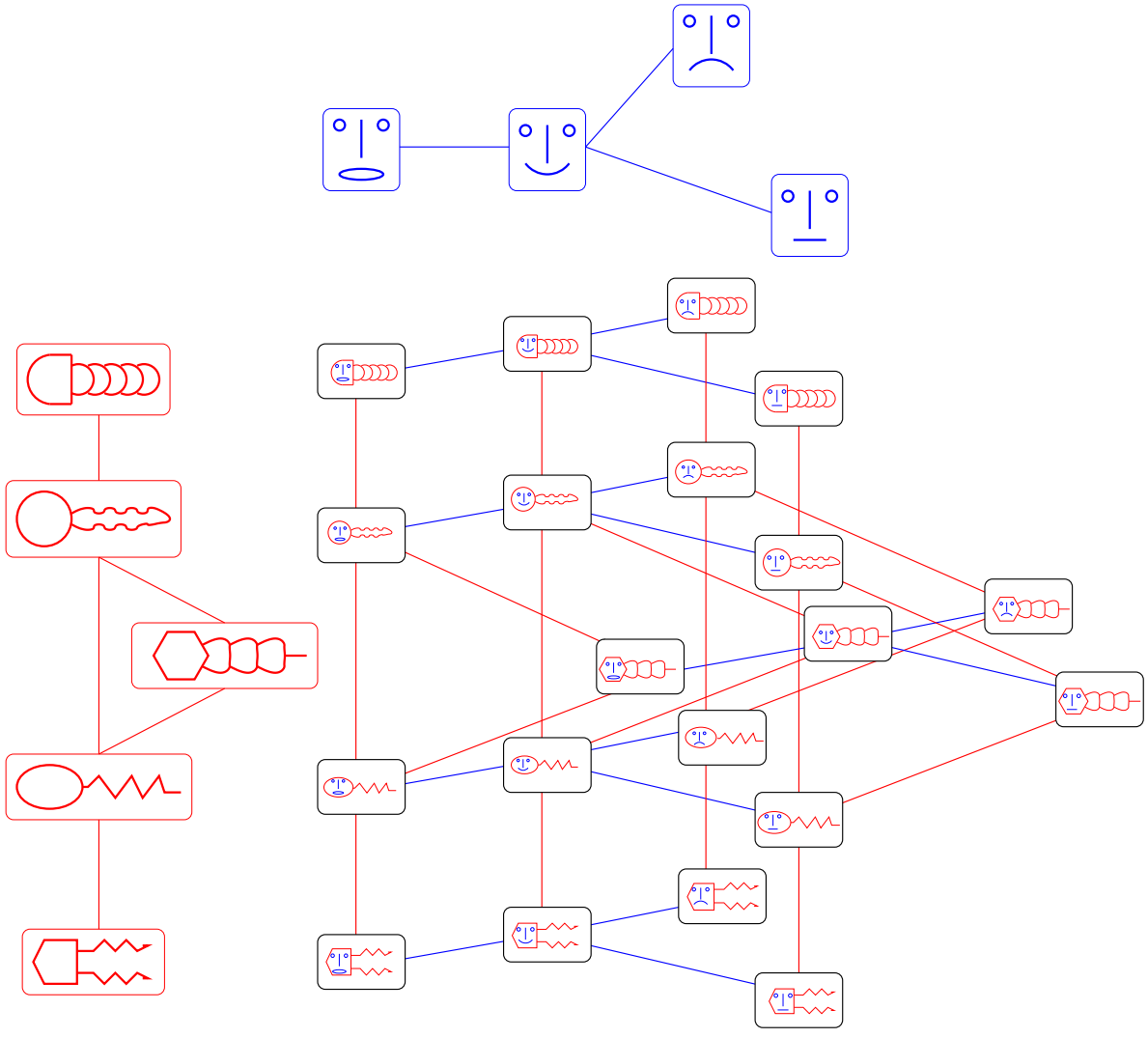
N is a neighborbood of (x_1, x_2) iff there are neighborhoods $N_1 \in \mathcal{N}_1(x_1)$ and $N_2 \in \mathcal{N}_2(x_2)$ such that $N_1 \times N_2 \subseteq N$.

The projections $\pi_i : (X_1 \times X_2, \mathcal{N}) \to (X_i, \mathcal{N}_i) : \ (x_1, x_2) \mapsto x_i$ are continuous functions for $i = 1, 2$.

(as in topological spaces)

What is a Phenotypic Character (Merkmal)?





Characters

Idea:

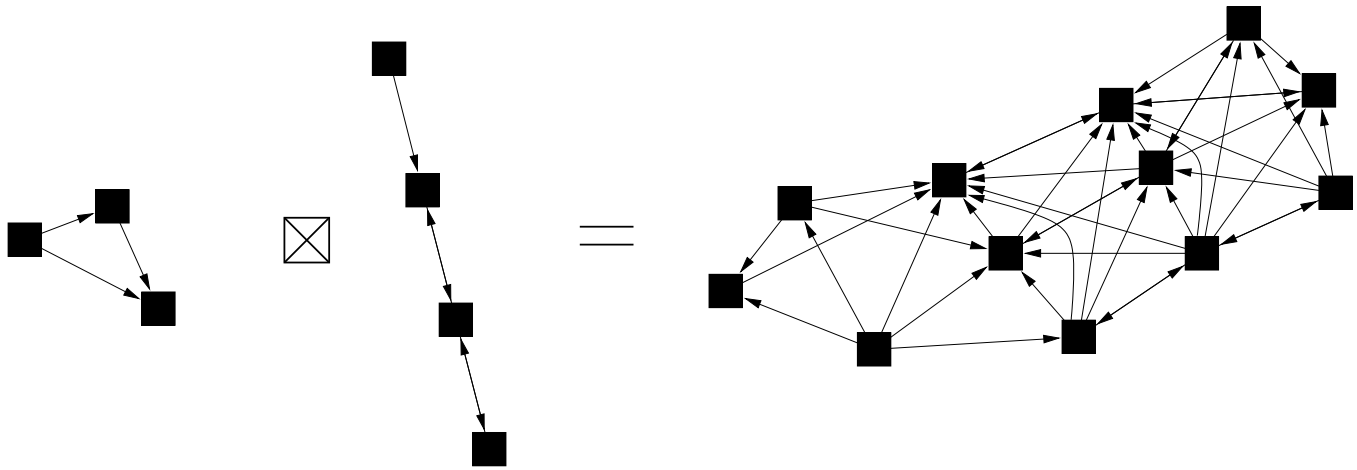
Characters **can** vary independently

$\Longleftrightarrow$

Factors of phenotype space

Mutation only: Directed Graphs

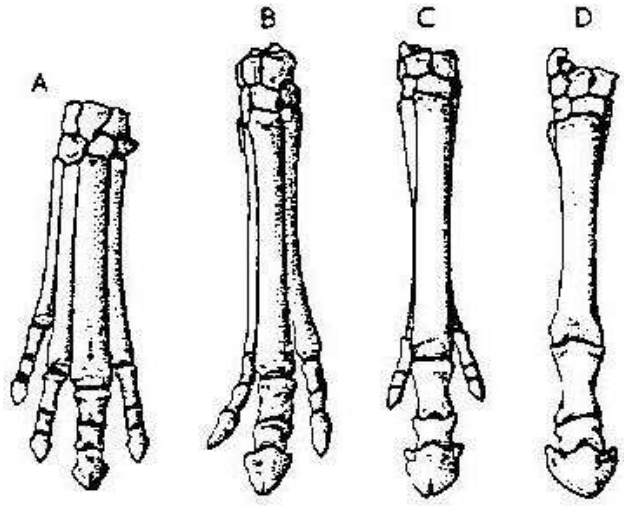
topological product $\Longleftrightarrow$ strong product of graphs



Unique prime factor decomposition of connected graphs and digraphs.

Allows identification of global characters.

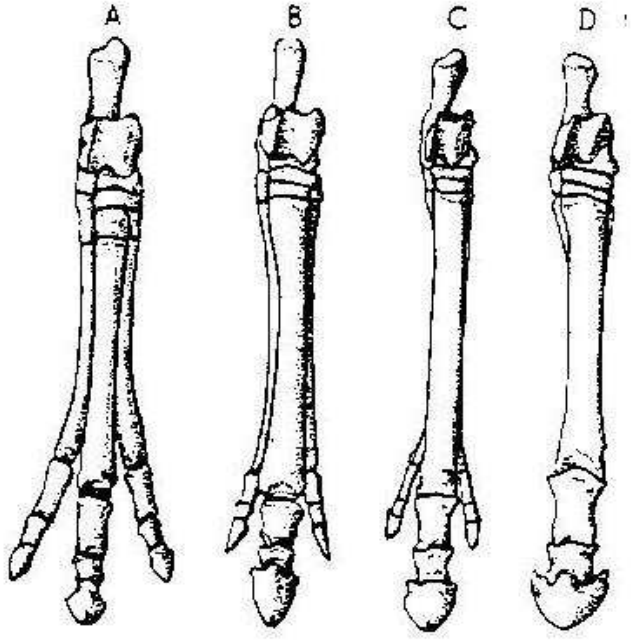
Homologous Characters



Horse Leg Evolution

A Hyracotherium
B Miohippus
C Merychippus
D Equus

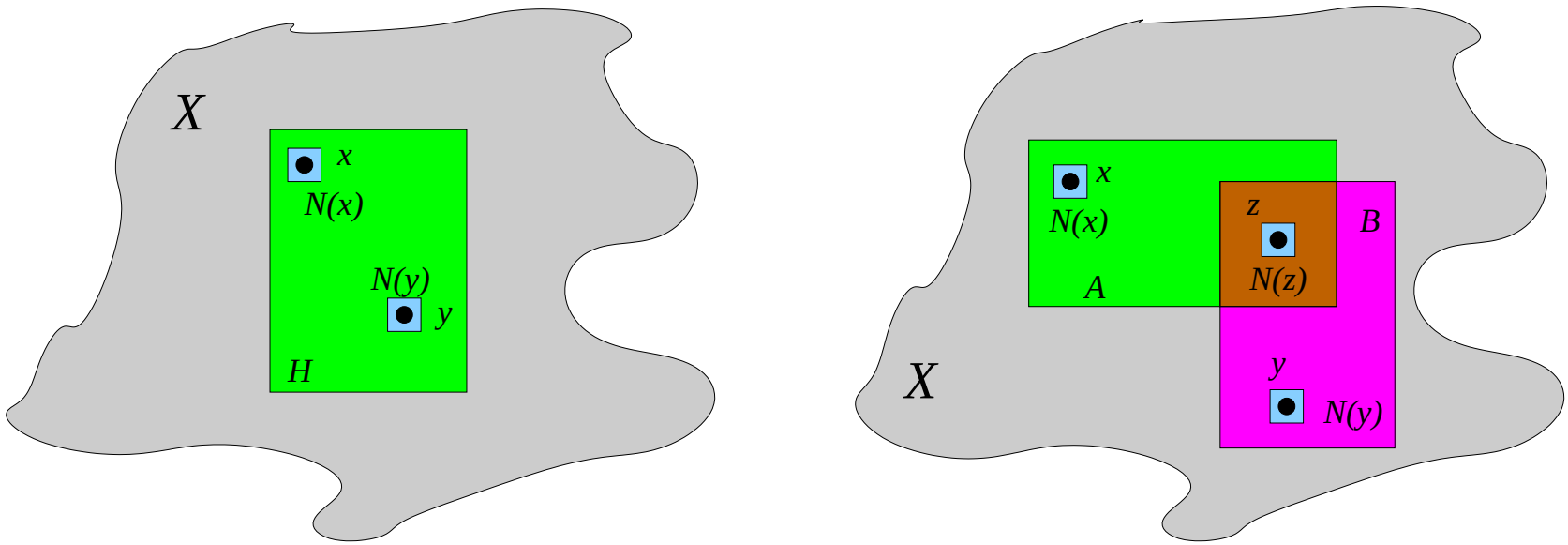
Forefeet



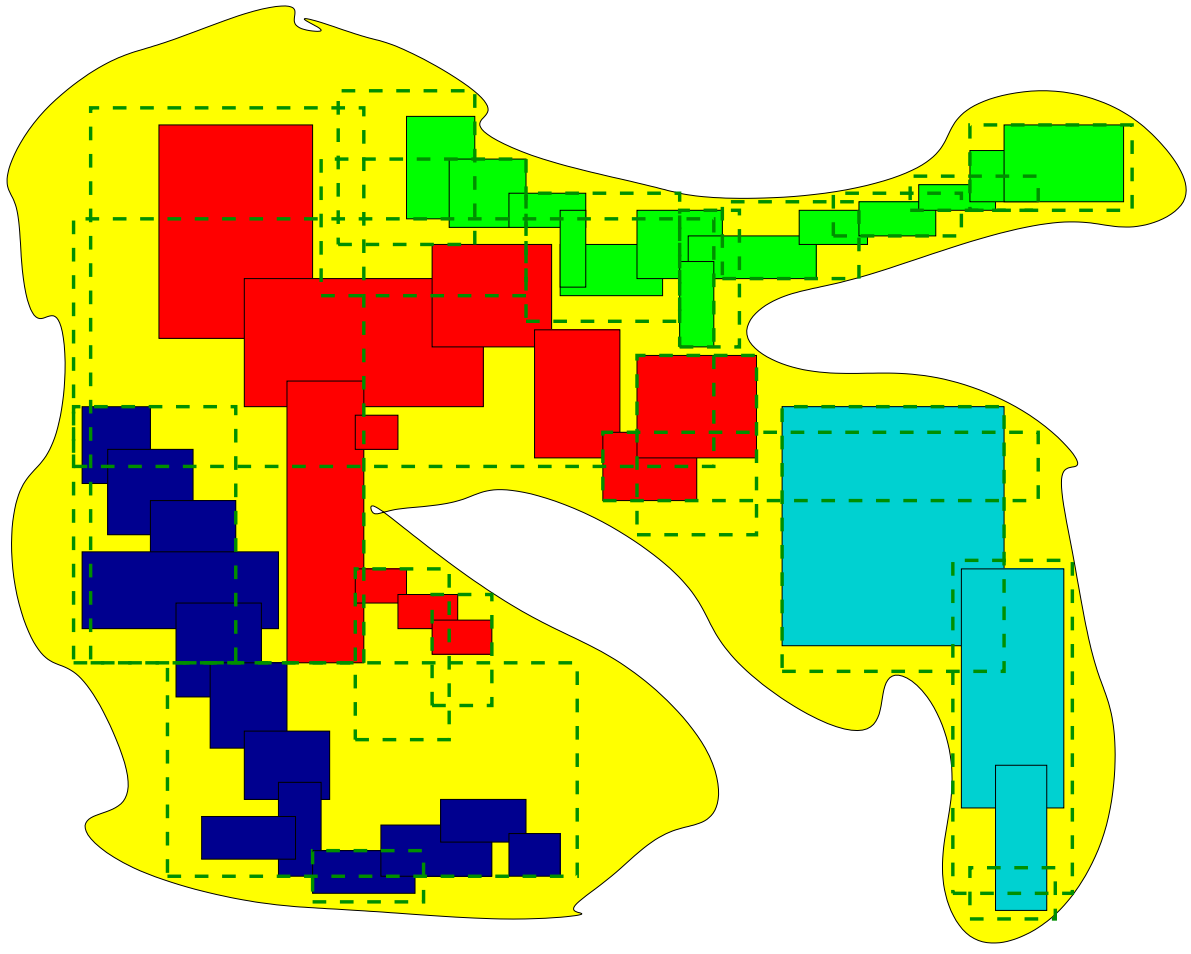
Hindfeet

Homology

If characters "are" just factors of the phenotype space, how do we know which factors are "the same" in different places of phenotype space??



A factorizable region H established the identity of characters between two points $x, y \in \text{int}(H)$. In a second step overlapping factorizable regions A and B can mediate character identity via points in their common interior $z \in \text{int}(A) \cap \text{int}(B)$.



The identity of characters can be extended wherever the colored rectangles overlap. The four types of factorizations (different colors) all may have one “highlevel”-factor (character) in common (green dashed outline) which can be made up of different combinations of “sub-characters”.

Summary on Molecular Evolution

We have developed here a **framework** (or a **language**) for formalizing evolution at large scales that can deal with:

- Continuous and discontinuous evolutionary transitions
- the concept of a **character**
- the concept of **homology**
- different notions of **innovation**
- & suggests (at least some) testable hypotheses

Of course, it is only a first step ...

Fitness Landscapes (again ...)

Fitness Landscapes: Optimization by Local Search

Given:

- A set X of configurations
- A set Z of (partially) ordered values
- A fitness function $f : X \to Z$
- Some collection $\mathfrak{M}$ of rules that transform subsets of X into other subsets of X. (For technical convenience we assume that $\mathfrak{M}$ contains an identity operator that simply leaves the subset unchanged.)

So: all structure on X is implied *somehow* by $\mathfrak{M}$.

A closure function

How does $\mathfrak{M}$ "explore" / "search" X?
Define $c : \mathcal{P}(X) \to \mathcal{P}(X)$ such that for each subset $A \subseteq X$ holds:
$c(A) =$
$\{x \in X | x$ can be generated instantaneously from (a subset of) A by means of $\mathfrak{M}\}$
"instantaneous" means a **single** application of an operator from $\mathfrak{M}$
Properties of c:

- $c(\emptyset) = \emptyset$
- $A \subseteq A'$ implies $c(A) \subseteq c(A')$
- $A \subseteq c(A)$

So (X, c) is a neighborhood space. This is nice.

Local Minima

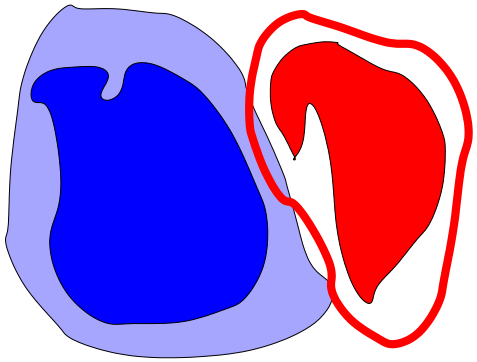
x is a local minimum of $f : X \to Z$ if there is a neighborhood $N \in \mathcal{N}(x)$ such that there is no $y \in N$ with $y < x$.

If Z is a po-set instead of ordered, these are the local Pareto points!

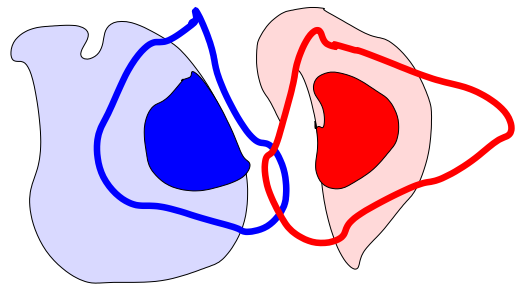
Maybe a stricter definition is desirable sometimes:
x is a local minimum of $f : X \to Z$ if every neighborhood of x contains a neighborhood $N \in \mathcal{N}(x)$ such that there is no $y \in N$ with $y < x$.

Connectedness in Neighborhood Spaces

"Separation" defines what is NOT connected.



semi−separated sets



equivalently: all pairs of subsets are semiseparated

Additional requirement for separation: If A and B are "separated" then combining subsets $A' \subseteq A$ and $B' \subseteq B$ does not generate anything novel, i.e.: $c(A' \cup B') = c(A') \cup c(B')$.

Connectedness in Neighborhood Spaces

Definition. Two sets A and B are productively separated if for any subset $U \subseteq A \cup B$ holds:

- $c(U \cap A) \cap B = \emptyset$,
- $c(U \cap B) \cap A = \emptyset$, and
- $c(U) = c(U \cap A) \cup c(U \cap B)$.

Definition. A subset A of X is *connected* if it cannot be decomposed into two non-empty productively separated sets.

Hawai‘i Theorem

Theorem. Connected sets in neighborhood spaces have the "usual" properties.

Hawai'i Theorem

- Individual points $\{x\}$ are always connected.
- If U' and U'' are connected and $U' \cap U'' \neq \emptyset$, then $U' \cup U''$ is connected
- If U is connected then $c(U)$ is connected.
- Let $\{U_\imath\}$ be a collection of connected sets with $x \in U_\imath$ for all $\imath$, then $\bigcup_\imath U_\imath$ is also connected.

Thus: For each set $A \subseteq X$ and each $x \in A$ the connected component

$$A[x] = \bigcup \{A' \subseteq A | x \in A' \text{ and } A' \text{ connected}\} \tag{1}$$

is well-defined.

Level Sets

For each $\eta \in Z$ define the level sets

$$\Lambda_\eta = \{z \in X | f(z) < \eta\}$$

For each $x \in X$ and each $\eta in Z$ define $\Lambda^*_\eta(x)$ as the connected component containing x, provided $f(x) < \eta$.
We set $\Lambda^*_\eta(x) = \{x\}$ if $f(x) = \eta$ and $\Lambda^*_\eta(x) = \emptyset$ otherwise.

$$\eta[x, y] = \inf\{\eta \in Z | \Lambda^*_\eta(x) \cap \Lambda^*_\eta(y) \neq \emptyset\}$$

Then x and y belong to the same connected component of the levelset at heights above $\eta[x, y]$ and to different components below $\eta[x, y]$.
THUS: $\eta[x, y]$ is the height of the saddle between x and y.
So what are saddle points?
... we need an appropriate notion of compactness to talk about this ...

So What?

- We can think about topological/geometric features of landscapes without making the distinction between discrete, continuum, low- or high dimensional
- We don't need symmetries of the move set, additivity (i.e., a graph structure), or a distance between points.
- It seems that the language for the most part does not care if we use integers, reals, vectors, or any arbitrary poset as value set.
- We can define barrier trees completely naturally for arbitrary problems with ARBITRARY genetic operators
- It was fun working on it.

And in Chemistry?

Peter Dittrich (U. Jena) constructed a theory of *Chemical Organizations* based on Walter Fontana's the notions of a

- "closed" sets of molecules
- "self-maintaining" set of molecules

Chemical Reactions

X . . . set of molecular types
$\mathfrak{R}$ . . . set of reactions:

$$\sum_x \bar{\alpha}_{x,\rho} \to \sum_x \alpha_{x,\rho}$$

Stochiometric coefficients: $\bar{\alpha}_{x,\rho}$ (l.h.s.) $\alpha_{x,\rho}$ (r.h.s.)
Domain and image of a reaction ρ:

$$\mathrm{dom}\rho = \{x \in X | \bar{\alpha}_{x,\rho} > 0\}$$
$$\mathrm{img}\rho = \{x \in X | \alpha_{x,\rho} > 0\}$$

Production Function

Consider a reaction mixture composed of molecules of types $A \subseteq X$.
The reaction ρ can take place in this medium if and only if $\mathrm{dom}\rho \subseteq A$.
Define the collection of all possible reactions that can start from A:

$$\mathfrak{R}_A = \{\rho \in \mathfrak{R} | \mathrm{dom}\rho \subseteq A\}$$

The products of all these reactions will have non-zero concentration after infinitesimal time.

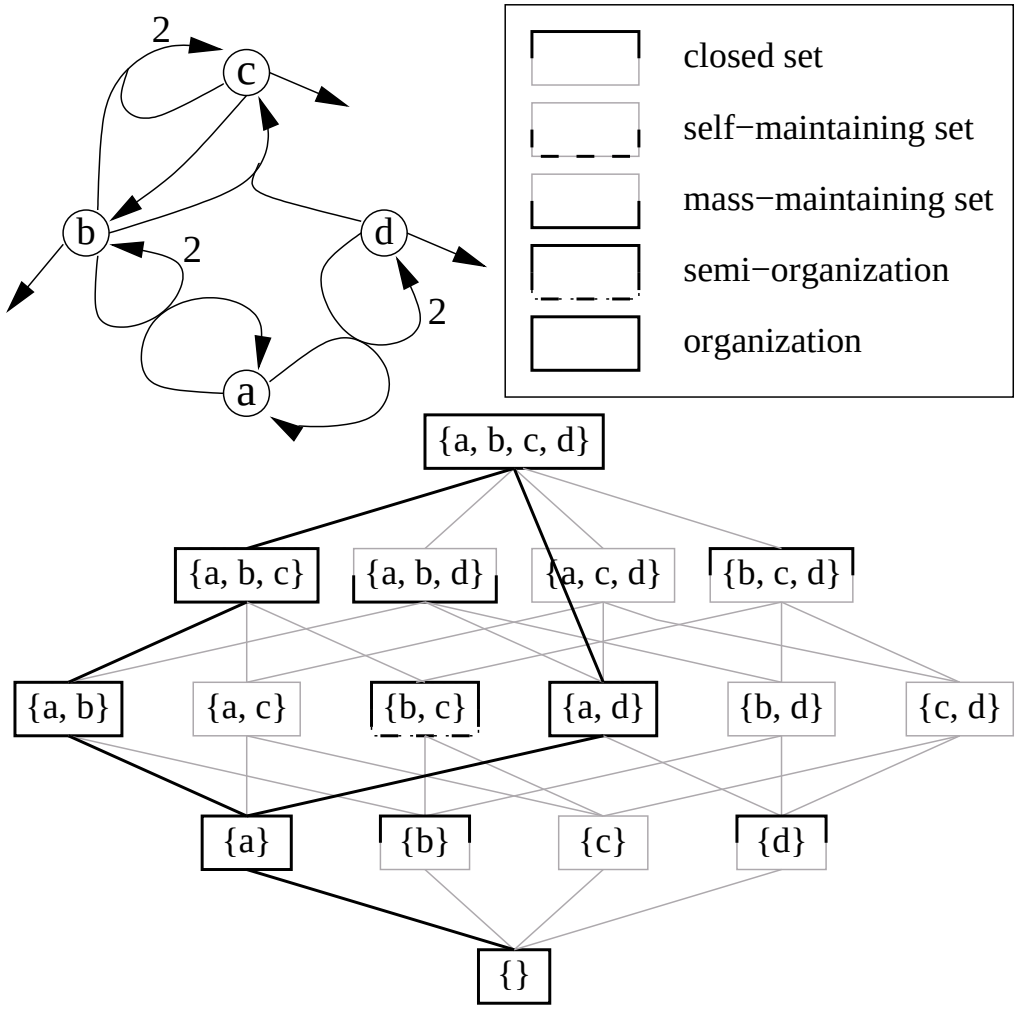
$$p(A) = \bigcup_{\rho:\mathrm{dom}\rho\subseteq A} \mathrm{img}\rho$$

We call $p : \mathcal{P}(X) \to \mathcal{P}(X)$ the production function of $(X, \mathcal{R})$.
Open question: To what extent is $(X, \mathcal{R})$ determined by p?

Refined Model

Stochiometric matrix of net reaction: $s_{x,\rho} = \alpha_{x,\rho} - \bar{\alpha}_{x,\rho}$
Stationary Fluxes J satisfy $\mathbf{S}J = 0$ and $J_\rho \geq 0$ (Clarke, Fell, . . .)
"outflux reactions" ψ_x for each species x
"influx reactions" ϕ_x for *some* species that are supplied from outside.
The flux vector J is said to produce x from A if $J_{\psi_x} > 0$ and $J_\rho > 0$ implies $\mathrm{dom}\rho \subseteq A$ for all non-outflux reactions ρ.
The species x is *maintainable* from A if there is a stationary flux vector J that produces x from A.
Let $s(A)$ denote the set of all species maintainable from A.
Then there is always a stationary flux vector that simultaneously produces all $x \in s(A)$ from A.
Theorem. $s(A) \subseteq p(A)$.
Furthermore, both $s(A)$ and $p(A)$ are isotonic.

Dittrich's Lattice of Organizations



P.Dittrich et al., archive/q-bio 0501016

Stationary state $\Rightarrow$ support is closed set w.r.t. s.

Chemical Organization Dictionary

The following works in the absence of inhibitory interactions:

Set functions	P. Dittrich's theory
$s(A) \subseteq p(A) \subseteq A$	"closed"
$A \subseteq p(A)$	self-maintaining
$A \subseteq s(A)$	mass-maintaining
$A = p(A)$	semi-organization
$A = s(A) = p(A)$	organization

THUS:Chemical organizations thus can be understood as topological constructs

Summary

- Large Chemical Reaction Networks can be understood as generalized topological spaces
- Chemical Organization Theory has a topological interpretation.
  Whether this correspondence is fruitful is a subject of current research
- Many basic results and constructions from classical point set theory still work for generalized closure spaces.
- Agende: explore whether other topological concepts such as denseness, connectedness, convergence, separation can be interpreted in chemical terms. After all, these concepts come for free with the topological interpretation of $(X, \mathfrak{R})$.

Acknowledgements

- Bärbel Stadler
- Günter Wagner and Max Shpak (Yale), Walter Fontana (Harvard) [molecular evolution]
- Gil Benkö, Christoph Flamm [simulation of large networks]
- Peter Dittrich (Jena) [lots of discussions on chemical organizations]