Fantasy, Brain Damage, and My Dreams

of Perfectly Simulating Genome Evolution

Roman Stocsits

Bled, 23rd February 2005

Overview

Overview

Overview

Overview

- Introduction
  - The Initiation of the Idea
  - Features of Genomes
  - Intrinsic Problems of Simulations
- The Plan
  - Existing Programs
  - The Plan of the New Algorithm
- Questions to be Asked
- Expected Problems
- Dreams and Science Fiction

Overview

- Introduction
  - The Initiation of the Idea
  - Features of Genomes
  - Intrinsic Problems of Simulations
- The Plan
  - Existing Programs
  - The Plan of the New Algorithm
- Questions to be Asked
- Expected Problems
- Dreams and Science Fiction

Overview

Overview

- Introduction
  - The Initiation of the Idea
  - Features of Genomes
  - Intrinsic Problems of Simulations
- The Plan
  - Existing Programs
  - The Plan of the New Algorithm
- Questions to be Asked
- Expected Problems
- Dreams and Science Fiction

Overview

- Introduction
  - The Initiation of the Idea
  - Features of Genomes
  - Intrinsic Problems of Simulations
- The Plan
  - Existing Programs
  - The Plan of the New Algorithm
- Questions to be Asked
- Expected Problems
- Dreams and Science Fiction

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

Polish wooden wodka and the initiation of ideas.

- A piece of wood makes the wodka slightly yellow and tasty.
- This is the main difference to Swedish and Russian versions.
- Those versions are only useful for cleaning and desinfection purposes.

What about 'Flaki' (?), and drinking alcohol against pain?

And what else?

Once upon a time in Poland...

At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

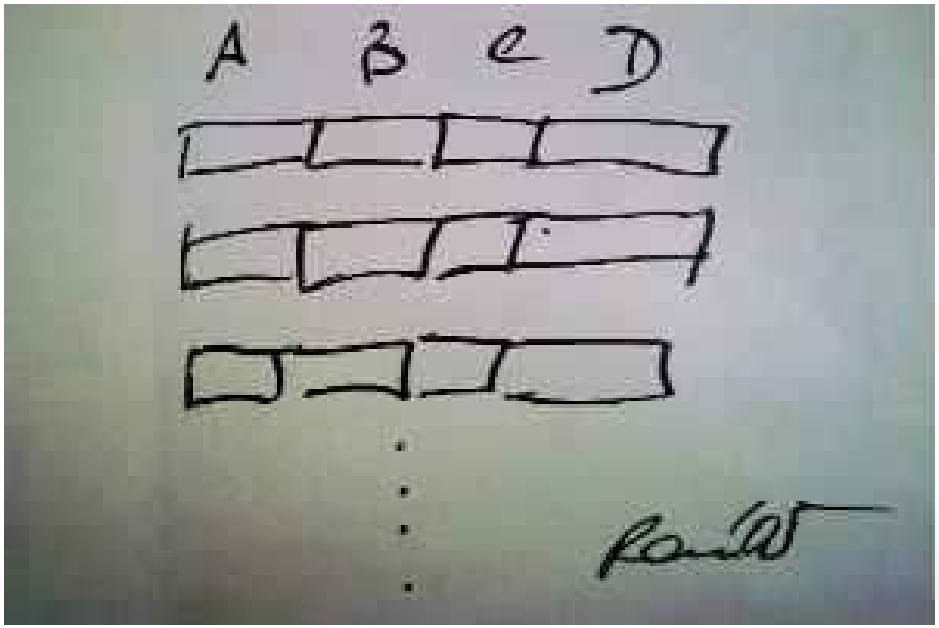
At the Institute for Microbiology in Wroclaw (Stanislaw Cebrat):

Simulations of genome evolution

- Model genomes are defined as arrays of protein coding genes.
- Randomly placed mutations influence fitness of encoded phenotype.
- The phenotype is defined as a set of (bio)chemical features.
- Artificial selective pressure acts on this phenotype.
- Iterations of mutation and selection

Once upon a time in Poland...

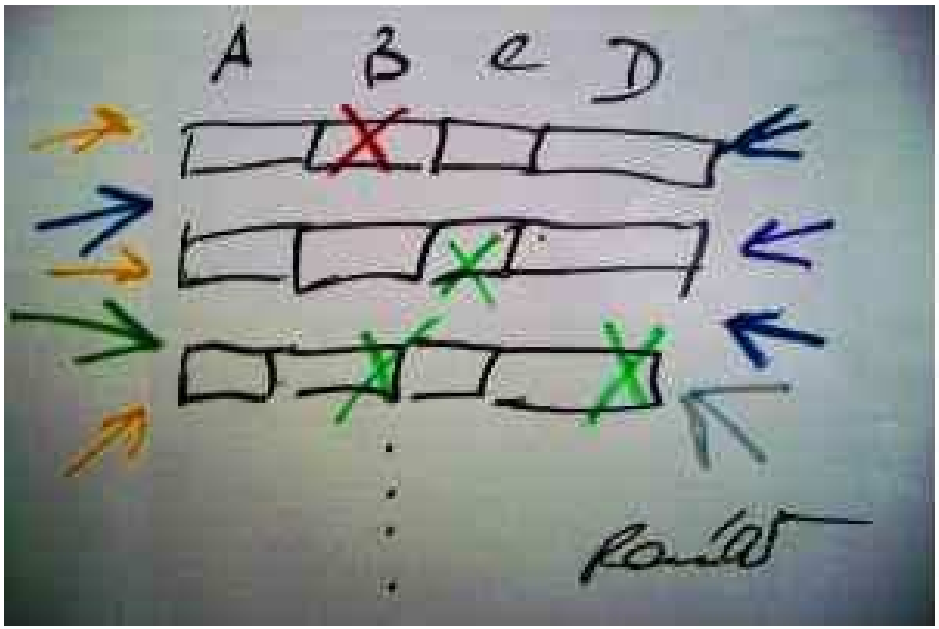
The procedure looks something like this (I hate `xfig`):



Starting from some 'genomes' with the same set of genes...

Once upon a time in Poland...

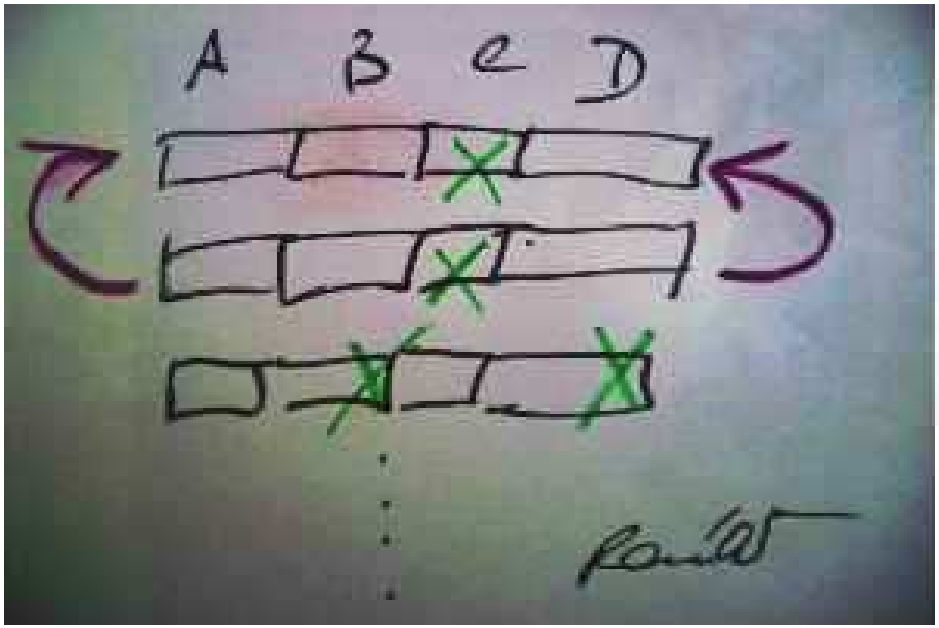
...mutations are inserted (about 1 per generation cycle).



Selective pressure acts, and some mutations are letal (red)...

Once upon a time in Poland...

The genome containing the letal mutation dies...



...and gets replaced randomly by another genome in the pool.

Once upon a time in Poland...

The phenotype might, for instance, consist of:

- isoelectric points
- amino acid sequence motifs
- secondary structure motifs

and also various others...

Fitness and Surviving in Poland

The decision if selection is survived
regarding a specific phenotypic marker (= gene)
might be just YES or NO.

But there might also be in-between states,
and fitness might even be continuously decreasing from
perfectly fitting to letal.

This means that a disadvantageous mutation is not necessarily letal
(*in vivo* and *in silico*).

Fitness and Surviving in Poland

The decision if selection is survived
regarding a specific phenotypic marker (= gene)
might be just YES or NO.

But there might also be in-between states,
and fitness might even be continuously decreasing from
perfectly fitting to letal.

This means that a disadvantageous mutation is not necessarily letal
(*in vivo* and *in silico*).

Fitness and Surviving in Poland

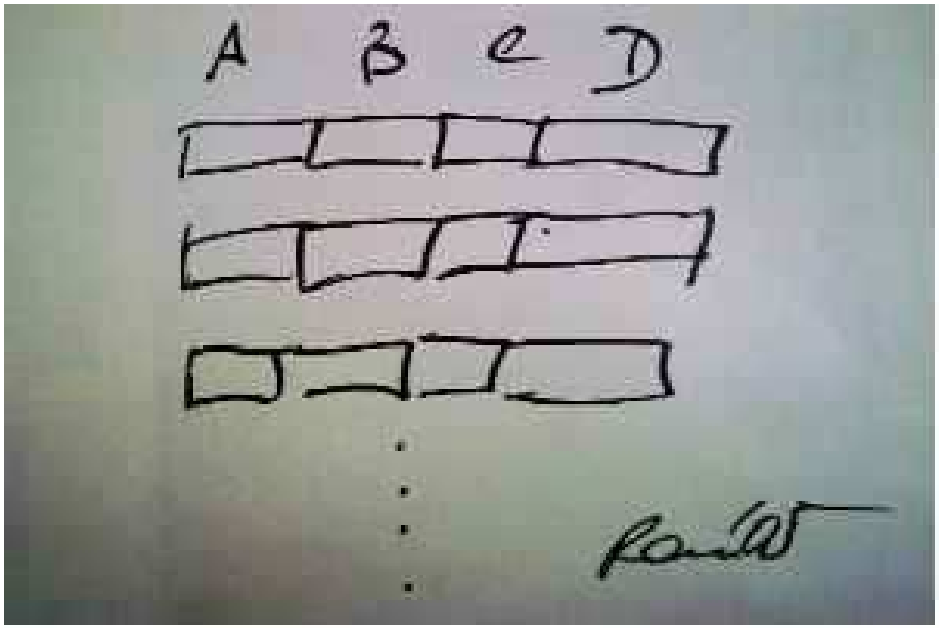
The decision if selection is survived
regarding a specific phenotypic marker (= gene)
might be just YES or NO.

But there might also be in-between states,
and fitness might even be continuously decreasing from
perfectly fitting to letal.

This means that a disadvantageous mutation is not necessarily letal
(*in vivo* and *in silico*).

Shortcomings, so far, are obvious

Nevertheless, a **real** set of genomes does not look like that...



At least, we think so...

Shortcomings, so far, are obvious

We can state:

- 'Flaki' (?) make it necessary to drink alcohol against pain.
- And genome simulation needs a model extended to functional RNA genes, various regulation sites and much, much more ...

Small 'Brötchens': The next step is

Extension to RNA genes — tRNA at the very beginning...

Shortcomings, so far, are obvious

We can state:

- 'Flaki' (?) make it necessary to drink alcohol against pain.
- And genome simulation needs a model extended to functional RNA genes, various regulation sites and much, much more ...

Small 'Brötchens': The next step is

Extension to RNA genes — tRNA at the very beginning...

Shortcomings, so far, are obvious

We can state:

- 'Flaki' (?) make it necessary to drink alcohol against pain.
- And genome simulation needs a model extended to functional RNA genes, various regulation sites and much, much more ...

Small 'Brötchens': The next step is

Extension to RNA genes — tRNA at the very beginning...

Yes, we know the world!

Some parts needed are **existing**, some are **under construction**, but some are still science fiction...

The Vienna RNA package

- features some essential routines we plan to use...
- ...and can be adapted to fit some eventual other needs.

Yes, we know the world!

Some parts needed are **existing**, some are **under construction**, but some are still science fiction...

The Vienna RNA package

- features some essential routines we plan to use...
- ...and can be adapted to fit some eventual other needs.

Features of Genomes

Genomes consist of genes (truly!)

Expression via complex regulation leads to a phenotype...

Phenotypes are "built up" by proteins, (lots of...) RNAs, weird complexes of all and everything, up to cells, tissues etc.

Function follows form (the opposite of modern industrial design...).

Genomes evolve via mutation and recombination...

...of variated defaults to get even more and faster variation of variation of variation etc.

"Das ist alles sehr kompliziert." (Fred Sinowatz, Burgenland)

Features of Genomes

Regarding regulation:

Cross reactions, (antisense) inhibitions, co-activation, reaction networks, and much more maybe widely unknown dependencies in 3 dimensions within the cell.

Enhancers, promoters, transcription factor binding sites cooperate with various types of regulatory sequence motifs...

Interactions among proteins, RNA-protein complexes, RNAs, DNA-protein complexes, up to the sub-organellar state...

Rapidly interchanging states of all the things above...

"Das ist alles sehr kompliziert." (Fred Sinowatz, Burgenland)

Intrinsic Problems of Genome Simulations

Complex processing of things that often are only poorly understood

- Complex processing of huge data amounts is of course expensive in time and memory.
  - With more realistic models: regulation is highly variable.
  - Lots of principles and mechanisms are widely not understood.
- Little is known about effective dependencies among various regulation schemata (regarding quality and quantity).

Let's go into some detail...

Intrinsic Problems of Genome Simulations

Complex processing of things that often are only poorly understood

- Complex processing of huge data amounts is of course expensive in time and memory.
  - With more realistic models: regulation is highly variable.
  - Lots of principles and mechanisms are widely not understood.
- Little is known about effective dependencies among various regulation schemata (regarding quality and quantity).

Let's go into some detail...

Intrinsic Problems of Genome Simulations

Complex processing of things that often are only poorly understood

- Complex processing of huge data amounts is of course expensive in time and memory.
  - With more realistic models: regulation is highly variable.
  - Lots of principles and mechanisms are widely not understood.
- Little is known about effective dependencies among various regulation schemata (regarding quality and quantity).

Let's go into some detail...

Intrinsic Problems of Genome Simulations

Complex processing of things that often are only poorly understood

- Complex processing of huge data amounts is of course expensive in time and memory.
  - With more realistic models: regulation is highly variable.
  - Lots of principles and mechanisms are widely not understood.
- Little is known about effective dependencies among various regulation schemata (regarding quality and quantity).

Let's go into some detail...

Intrinsic Problems of Genome Simulations

Complex processing of things that often are only poorly understood

- Complex processing of huge data amounts is of course expensive in time and memory.
  - With more realistic models: regulation is highly variable.
  - Lots of principles and mechanisms are widely not understood.
- Little is known about effective dependencies among various regulation schemata (regarding quality and quantity).

Let's go into some detail...

Intrinsic Problems of Genome Simulations

At a first glance, genomes are 'only' very long strings of letters.

But, as always in evolution, **selective pressure acts on function**.

And genomes function indeed in 3 dimensional space. Together with vast amounts of co-operators.

Thus, the genome definition can be extended to the whole system of encoded data AND functional **structures** (protein, RNA, DNA (sic!), the complete cytoskeleton ...).

Intrinsic Problems of Simulations

The plan is long time genome simulation, and it seems to be impossible...

When genome evolution (selective pressure on genomes) is in question: Genomes cannot be reduced only to template DNA MIND: function is the decisive part for selection.

The regulation of the template is, at least, as decisive as the contents of the template.

Therefore, we cannot make a difference between regulation and contents, if we look at selective pressure on function and the consequences for evolution, in the case of genomes.

Intrinsic Problems of Simulations

The plan is long time genome simulation, and it seems to be impossible...

When genome evolution (selective pressure on genomes) is in question: Genomes cannot be reduced only to template DNA

MIND: function is the decisive part for selection.

The regulation of the template is, at least, as decisive as the contents of the template.

Therefore, we cannot make a difference between regulation and contents, if we look at selective pressure on function and the consequences for evolution, in the case of genomes.

Intrinsic Problems of Simulations

The plan is long time genome simulation, and it seems to be impossible...

When genome evolution (selective pressure on genomes) is in question: Genomes cannot be reduced only to template DNA MIND: function is the decisive part for selection.
The regulation of the template is, at least, as decisive as the contents of the template.

Therefore, we cannot make a difference between regulation and contents, if we look at selective pressure on function and the consequences for evolution, in the case of genomes.

Intrinsic Problems of Simulations

The plan is long time genome simulation, and it seems to be impossible...

A "phenotype of a genome" is its scheme of regulation, recombination and expression **first**.

And only as a consequence of this it is the way the animal, plant or bacterium looks like.

As the second step selective pressure acts on the animal, plant, fungus etc. in its environment.

First it acts on the genome functionality (correct regulation/expression/stability etc.)

Intrinsic Problems of Simulations

The plan is long time genome simulation, and it seems to be impossible...

Therefore, when defining a "genotype of a genome" we ought to speak especially about regulatory elements on DNA interacting with RNA/protein gene products.

These elements are not transcribed, but they function. — They are part of phenotype AND genotype.

It is sometimes not possible to strictly make a difference between phenotype and genotype.

On level of genome evolution the phenotype widely is the genotype.

Existing Programs and Partial Solutions

Existing Programs So Far:

- iterative mutation/selection algorithms (generation cycles)
- protein selection models
- RNAfold/RNAalifold
- Parallelization

Still Missing Parts of the Algorithm for Beginning

RNA selection models (only for tRNA at the beginning)

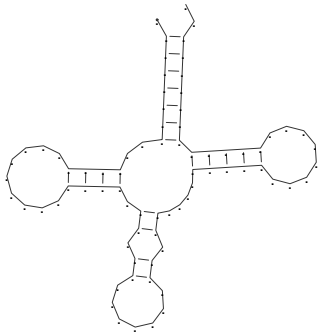
The first attempts for tRNA just feature a YES/NO selection.

For extending the existing algorithms to biologically more relevant genome simulations **step by step** it will be necessary to introduce in-between states.

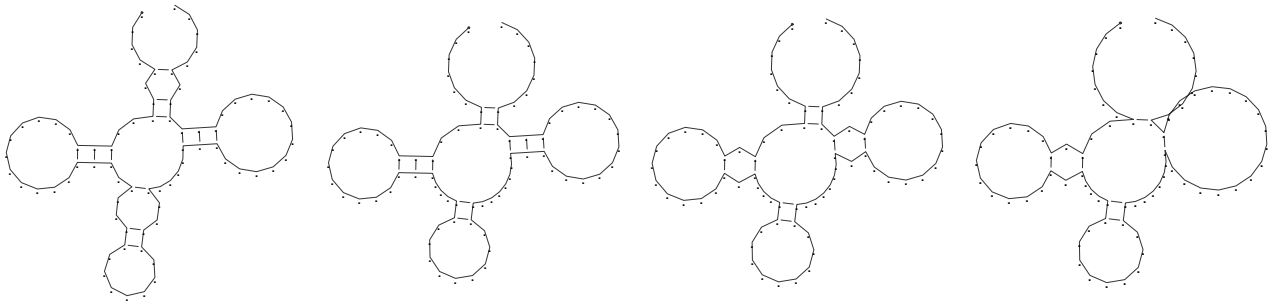
Routines still missing that catch, store and evaluate all important data about the behaviour of the system in long term simulations.

How do we get the tRNA selection models

We start from *real* tRNA sequences.



We allow certain variation from our default:



More or less stringent constraints must be still fulfilled after mutation for survival of the genome.

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

How do we adapt the tRNA selection model?

Using the existing...

We start the evolution simulation from *real* tRNA sequences.

- `RNAalifold` produces consensus structures
- We manually adapt the consensus to make it more or less stringent, depending on our needs (of course arbitrarily).
- At the beginning our artificial genome features really existing tRNA sequences
- The evolution simulation mutates the genomes and checks for fitness iteratively.
- Already existing parallelization software for genome evolving cycles on clusters

Evolving...

Variations are straight forward:
For extending to biologically more relevant genome simulations it is necessary to introduce in-between states.

e.g. if one constraint is not fulfilled, the survival is maybe in spite of that possible, if another constraint is fulfilled especially good.

Questions to be Asked...
...and Hopefully Answered

How behave various combinations of protein, tRNA, rRNA, ncRNA, regulatory DNA elements, and junk DNA...

Effects of junk DNA on selection?

YES if mutation-absorbing?
NO if genome length is not good for fitness?
(e.g. if generation time is an advantage...)

Expected Problems

How can we simulate improvements after newly invented features?
Protein: fitness via physical (?) parameters is contiuously varying.
BUT RNA:
fits into a SECOND more stringent consensus, not only the first less stringent consensus?

Dreams and Science Fiction

Applying to real genomes?
Do they evolve to something else also already existing...?
Related species...?
Prediction of mt-genome evolution?
Recurrencies regarding schemata, motifs, expression regulation networks...?

I am dreaming of a white Xtof

mitochondrial genomes -¿ reconstructing rearrangements and ALL ???

Prediction of future genomes (millions of years to come will show....)

Deducing Correlation between genome structure and evolutionary success?

Does genome structure influence speed of evolution?

etc., etc., etc....

I can't get no satisfaction...

This is **real** science fiction:

the perfectly realistic artficial genome *in silico* that explains all and everything.

(And maybe some time confirms '42' (Adams et al.))

I can't get no satisfaction...

This is **real** science fiction:

the perfectly realistic artficial genome *in silico* that explains all and everything.

(And maybe some time confirms '42' (Adams et al.))

Oh my God, you have been such a great audience!

Thank you...