CREST

Contamination of Computational Biology

Genetic improvement of computational biology software, WB Langdon and Karina Zile, Evolutionary Computation in Computational Biology, ECCSB 16 July 2017, José Santos *et al*., GECCO-2017 workshop. Room Topas1

W. B. Langdon

Department of Computer Science

CREST

Contamination of Computational Biology

W. B. Langdon
Department of Computer Science

Genetic Improvement of Computational Biology Software

- My thanks to ECCSB-2017 organisers
- Paper deals with 2 topics
  1. genetic improvement of computation biology software
  2. improving data used in computational biology
- Concentrate on 2. problems with data

Contamination of Computational Biology

- How we found mouldy data in the human genome
- An upside
- The reverse:

  human DNA in many species' reference genomes
- Problems of petabyte data cleaning
- Lessons

Mycoplasma Gene in the Human Reference Genome

- Visit to Matt Arno (King's College, London)
- Triplicate treatment v. control microarray
  - 1st 600 fold difference treatment v. control
  - 2nd 200 fold
  - 3rd nothing
- What is gene?
  - Affymetrix microarray HG-U133 +2 probeset 1570561_at was derived from GenBank AF241217
  - AF241217 "Homo sapiens unknown sequence" was submitted to GenBank in 2000

Run BLAST against everything

- Blast used to compare AF241217 DNA sequence with all sequenced species
- AF241217 sequence matches itself and various species of Mycoplasma

EBI > Tools > Sequence Similarity Searching > NCBI BLAST

NCBI BLAST Results

Summary Table | Tool Output | Visual Output | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	E()
☑ 1	EM_HTG:AF241217	Homo sapiens unknown sequence. *Cross-references and related information in:* ▶Ontologies	249	225	100.0	1.0E-121
☑ 2	EM_PRO:FJ876260	Mycoplasma orale strain MT-4 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Ontologies	2231	218	100.0	1.0E-117
☑ 3	EM_PRO:AF294995	Mycoplasma orale 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Literature ▶Ontologies	738	218	100.0	1.0E-117
☑ 4	EM_PRO:JN689375	Mycoplasma orale isolate LJH 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Ontologies	535	214	99.0	1.0E-115
☑ 5	EM_PRO:AY737010	Mycoplasma orale 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Literature ▶Ontologies	882	214	99.0	1.0E-115
☑ 6	EM_PRO:AY762640	Mycoplasma indiense 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence.	870	207	99.0	1.0E-111

NCBI Gene Expression Omnibus

- NCBI GEO is an archive of many thousands of gene expression datasets, including HG-133 +2.
- About 1% of published data contaminated with mycoplasma (33/2757) [BioTechniques, 47(6)]
- Positive response from authors of published papers.
- 2nd example found (and reported) [arXiv:1106.4192]
  - Uploaded to DNA data bank of Japan, DDBJ
  - Overnight to NCBI Washington, DC
  - Overnight to EBI, Cambridge

Mycoplasma Now known Problem

- Mycoplasma in 1000 Genomes Project [BioData Mining 7:3]
- Microbiologists recognise problem in wet labs.
- Do not recognise problem *in silico*

Contamination in other direction Human genes → other species

Human genes in hundreds of non-primate DNA sequence databases

- E.g. bacteria, plants and fish.
- Examples found in most phyla

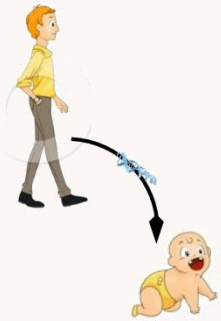
Big Data Clean up

- Problems have been repeatedly reported. Acknowledged but response was “we only curate data for the biologists” (I.e. those who uploaded crap, in 2000, have to fix it)
- Attitude changing?
- NCBI holding petabytes 10^{15} of data online
- Data cleaning/Data Wrangling is a well known problem
- Labour intensive
- Manual methods cannot scale!

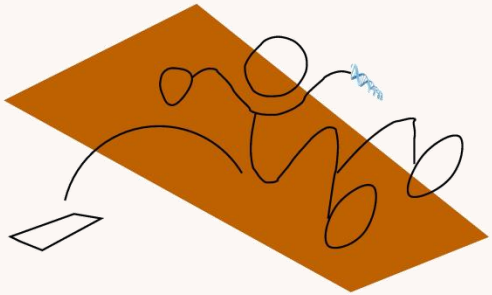
Genes Spread

- Microbes infect microbiology laboratories
- 2 genes have been copied into GeneBank
  - 1 via Japan, 1 into commercial tool. Others? patents?
  - Many human genes in nonprimate databases
- Data are routinely copied, allowing virtual genes (venes) to spread globally.
- Laboratories routinely sterilise glassware. They do not sterilise their computers.

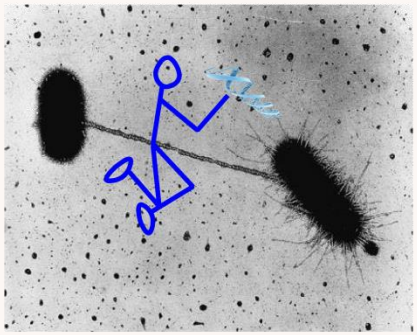
Genes Jump Silicon Barrier



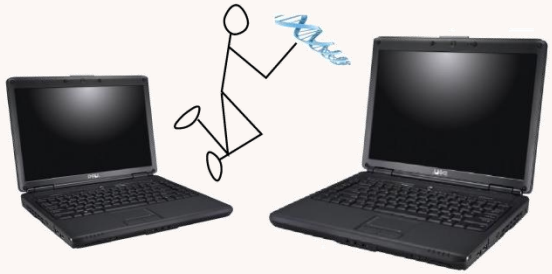
Mendel 1865



Jumping genes
McClintock 1930



Horizontal gene transfer
1959



Gene transfer to GenBank
Today

- 1865 vertical gene transfer
- 1930 gene transfer along chromosomes
- 1959 antibiotic resistance between species
- Jumping genes escape biology, cross the silicon barrier and roam computer databases

Conclusions

- Cultural disconnect
  - Computer scientists may assume Biological data is ok (Biologists know it isn't)
  - Biologists cannot conceive that their computer may be contaminated.
- Need to be skeptical of others' data (as we would be of our own).

END

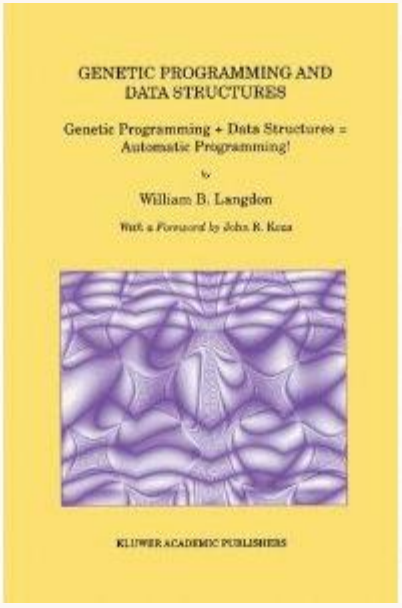
http://www.cs.ucl.ac.uk/staff/W.Langdon/ http://www.epsrc.ac.uk/ EPSRC

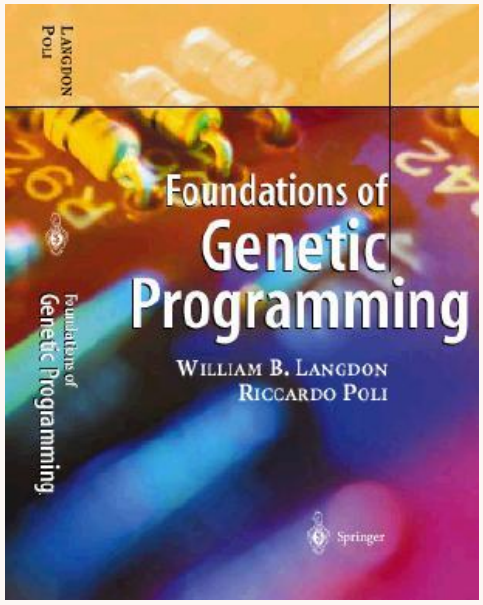
Genetic Improvement

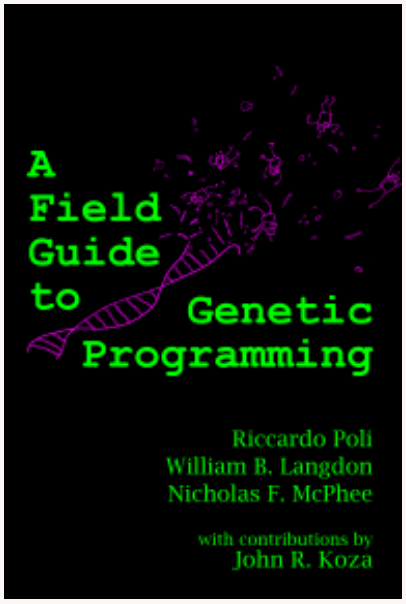
W. B. Langdon

CREST

Department of Computer Science







The Genetic Programming Bibliography

http://www.cs.bham.ac.uk/~wbl/biblio/

11628 references, 10000 authors

Make sure it has all of your papers!

E.g. email W.Langdon@cs.ucl.ac.uk or use | Add to It | web link

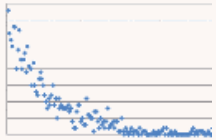
Co-authorships

RSS Support available through the Collection of CS Bibliographies.

Co-authorship community.
Downloads

Downloads by day

A personalised list of every author's GP publications.



Your papers

blog

Search the GP Bibliography at
http://liinwww.ira.uka.de/bibliography/Ai/genetic.programming.html