GECCO = GP + GA + ES + EP + EH + ER + DNA + LCS + RWA + AAAA +

GECCO 2004 Genetic and Evolutionary Computation Conference

June 26 - 30, 2004
(Saturday-Wednesday)
Seattle, Washington USA

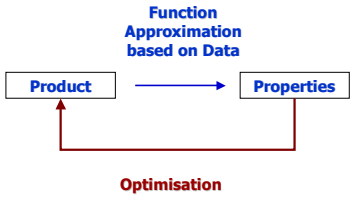
Evolutionary Product Design: Rocket Science or Art?

EVOLUTIONARY COMPUTATION IN INDUSTRY SESSION

Shail Patel
Unilever Corporate Research
Colworth House
Sharnbrook, Beds UK
shail.patel@unilever.com

Paradigm

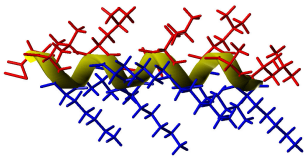
- Model & Optimise



>Case 1>Problem Description

Design an anti-bacterial peptide

- Natural Representation:
  - String of amino acids
  - **KLALALLLAGSAASA**

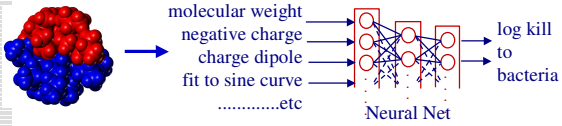


- Alpha-Helical Peptide
- Side View
  - red = hydrophobic
  - blue = hydrophilic

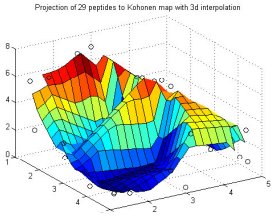
Patenting computer-designed peptides, Patel, Stott, Bhakoo & Elliott, Journal of Computer-Aided Molecular Design, 12: 543-556 1998

Fitness Function

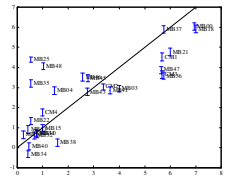
- Two Step Function

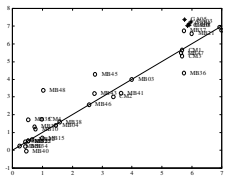


Non-linear Mapping

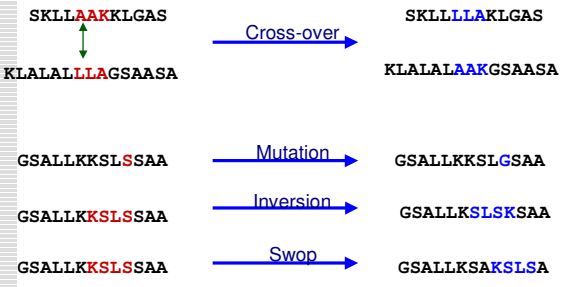


	xvalid R2 Linear	xvalid R2: Neural Net
Aureus	0.77	0.91
E Coli	0.75	0.81

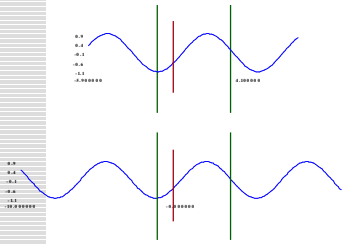




Operators



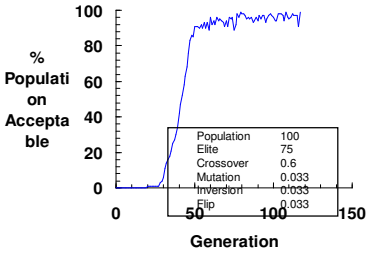
Special Cross-Over Operator



- Represent as right and left halves, separated by 'zero crossing' of hydrophobicity (red line in graph)
- Canonically align by right and left halves
- Select random (aligned) segments
- Perform cross-over

Running the GA

- Convergence in about 50 generations



Results: New Peptides

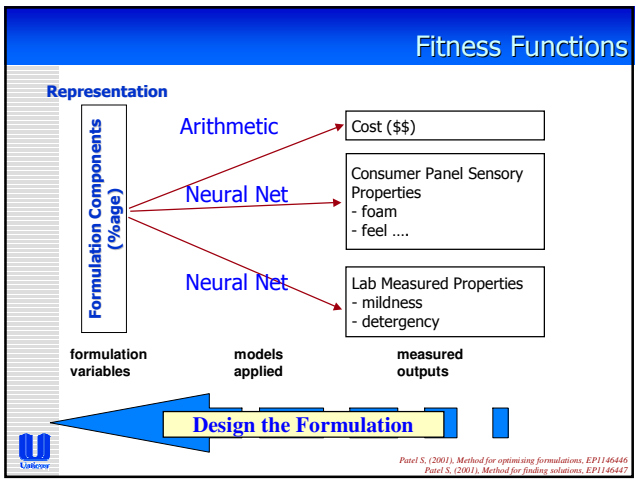
Peptide Sequence	predicted log kill	measured log kill
AASKAAKTLAKLLSSLLKLL	7.22	> 5.96
LLKKLLRAASKALSLL	7.13	> 5.90
AAKKLSKLLKTLLKLL	7.35	> 5.76
KALKKLLKLASSLLTAL	7.04	5.90
AASKALRTASRLARSLLTLL	7.03	> 5.85

>Case 2>Problem Description

Multi-Criteria Formulation Optimisation

- Current Process:
  - requirements
  - formulators experience
  - past products
  - experiments
- New Process:
  - requirements
  - formulators experience
  - past products
  - experiments
  - *predictions based on past experiments*
  - *trade-off analysis based on predictions*

Fitness Functions

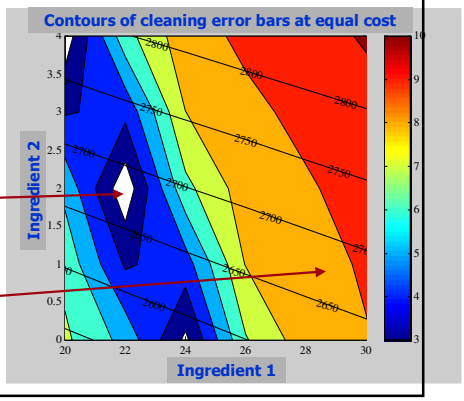


Patel S, (2001), Method for optimising formulations, EP1146446
Patel S, (2001), Method for finding solutions, EP1146447

Optimising Under Uncertainty

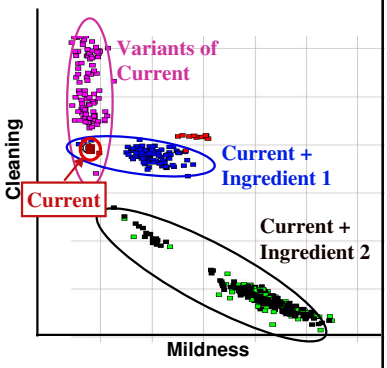
Same Cost
LOW Uncertainty

Same Cost
HIGH Uncertainty



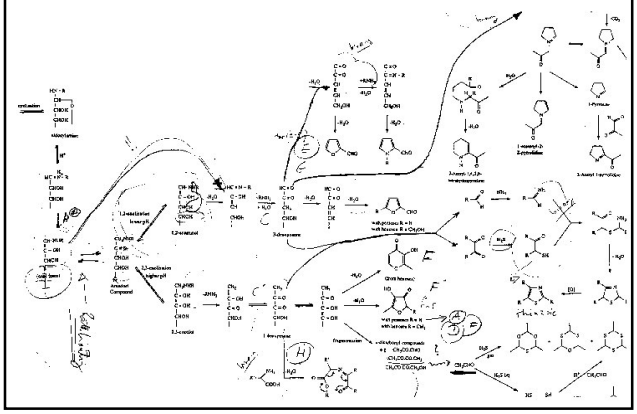
Results - Pareto Surface

- Optimisation against
  - Cleaning
  - Mildness
- BUT ALSO
  - Cost
  - Foam
  - Uncertainty
- Robustness ~= size of regions



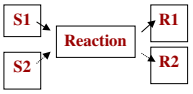
>Case 3>Problem Description

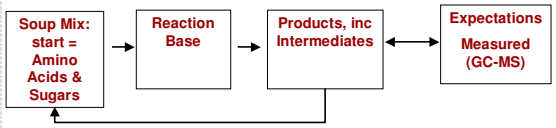
Modelling Maillard Food Browning Reactions



Iterated Reaction Graphs

- Model Reactions by input-output
- Model reactants & products / intermediates as list of molecules
- Iterate in Parallel
- Amend / Optimise Reaction Base
- Optimise Start Reactants





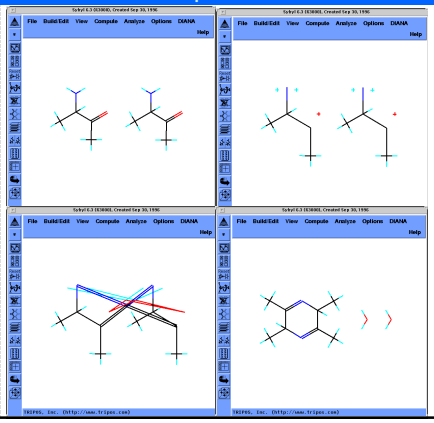
Patel S, Rabone R, Russell S, Tissen J, Klaffke W, (2001) Iterated Reaction Graphs: Simulating Complex Maillard Reaction Pathways, Journal of Chemical Information and Computer Sciences, Vol 41, 4, 926-933

Representation

- List of Molecules in SMILES notation

C(O)C(O)C(O)C(O)C(O)C=O
C(O)C(O)C(O)C(O)C(O)C=O — **Sugars**
C(O)C(O)C(O)C(O)C(O)C=O
C(O)C(O)C(O)C(O)C(O)C=O
…
NC(C(C)O)C(=O)O
NC(C(C)O)C(=O)O — **Amino Acids**
…….

'Operators' = Chemical Reaction



Chemical Reaction =

breaking bonds

making bonds

3-amino-2-butanone ↦ tetramethyl pyrazine precursor

Output

- Reaction Triplets := <Substrate Reaction Product>

C(=O)C(C(=CC(=C)O)O)O R1_1_6_enediol C(=C(C(=CC(=C)O)O)O)O 1
C(=O)C(C(C(C(=C)O)O)O)O R1_1_6_enediol C(=C(C(C(C(=C)O)O)O)O)O 1
C(=NCC(=O)O)C(C)O R1_1_2_enamine C(=C(C)O)NCC(=O)O 1
C1=CN=C(C)C(O)O1 R1_4_2_strecker C(=CN)OC(=O)C(C)O 1

- Chain to Form a Graph, eg one path:

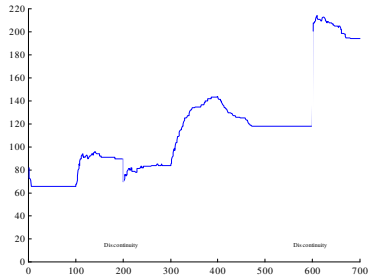
<Start> C(O)C(O)C(O)C(O)C(O)C=O R1_3_6_sub_enol C=C(O)C(O)C(O)C(O)C=O R1_1_8_abunsaturated C=C(O)C(O)C=C(O)C=O R1R_1_8_abunsaturated C=C(O)C(O)C(C(O)C=O)O R1_2_2_retroaldol C=C(O)C(O)C=O R2_1_1_Schiff_base CC(O)C(N=CC(C(=C)O)O)C(=O)O R1_4_1_Strecker CC(O)C=NC=C(C(=C)O)O R1_4_2_strecker CC(O)C=O R1_1_6_endiol ….

- List of products

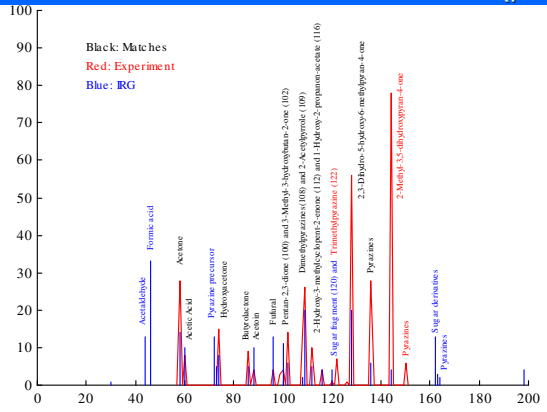
1 x C(=C(C)O)NCC(=O)O
1 x C(=CC(C(=O)OCC=N
14 x C(=C)O)O
1 x C(=O)C

- Virtual Mass Spec
  - (in time)
- Distance Measure
  - to expected products

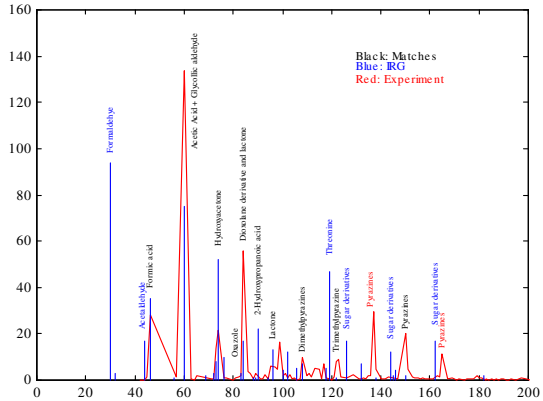
Glycollic Aldehyde fragment vs Time



Results for Glucose + Alanine (pH 7)



Results for Glucose + Threonine (pH 7)



Conclusion

- Model and Optimise paradigm is powerful
- Interplay between representation, operators and fitness function is critical
- Multi-criteria optimisation offers choices