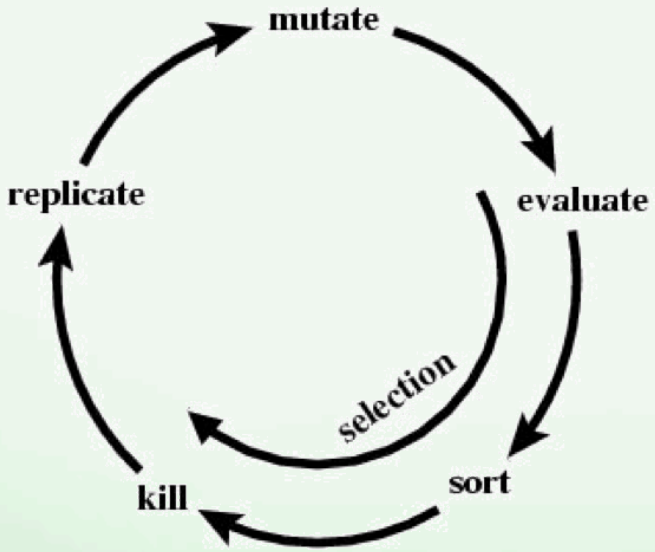
Multiple Sequence Alignment with Genetic Algorithms

Andrew Francl

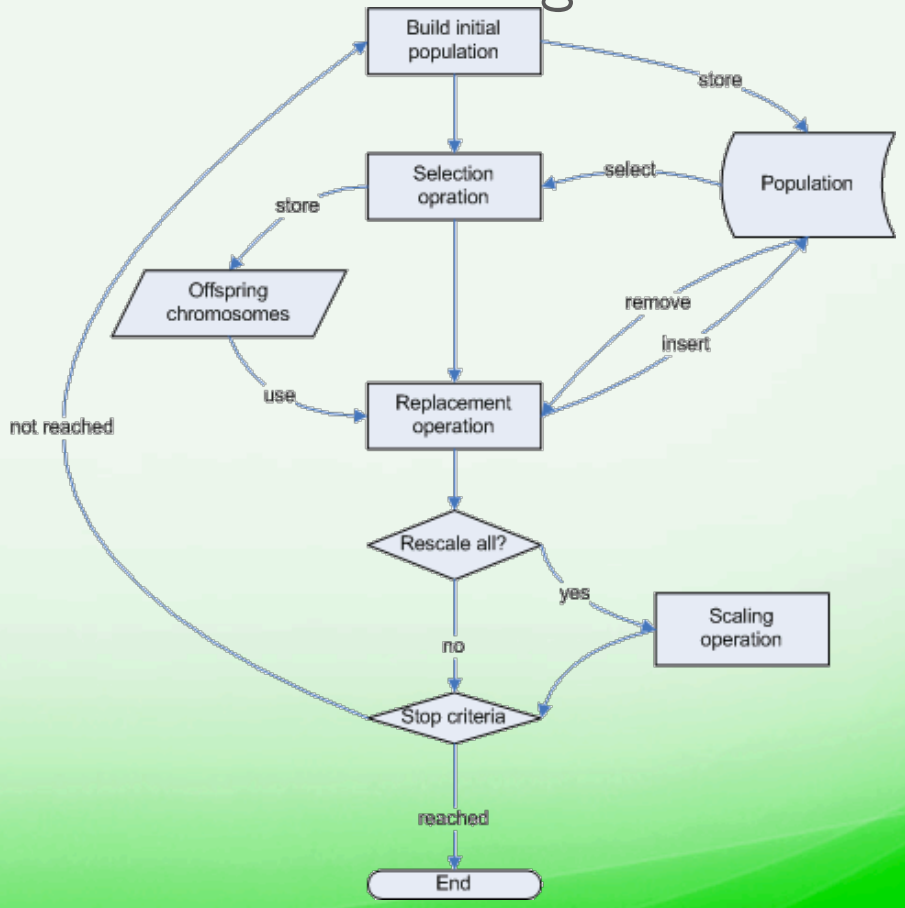
Introduction to Bioinformatics

Genetic Algorithms

Biological Evolution

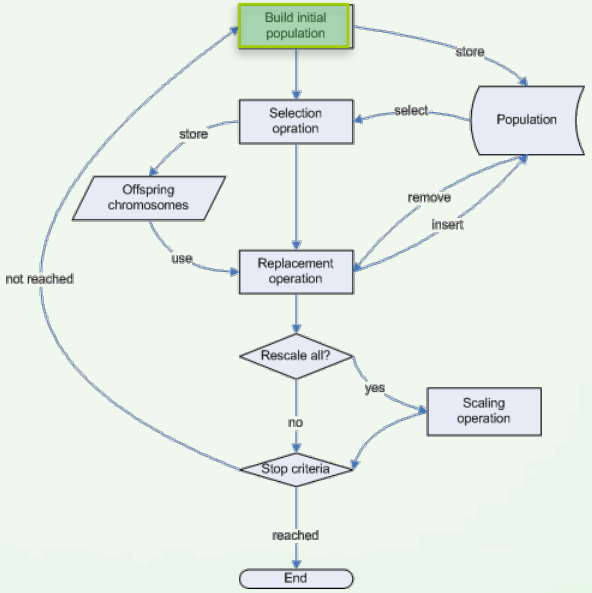


Genetic Algorithm



Evolution Process

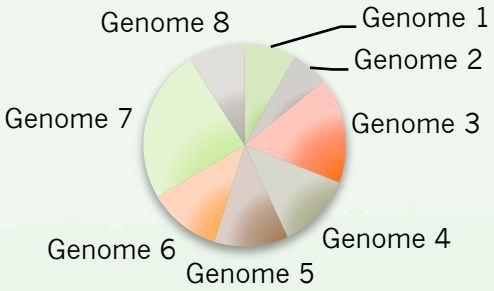
- Define initial parameters
- Encode Parameters into genome
- Create population of randomly created genomes



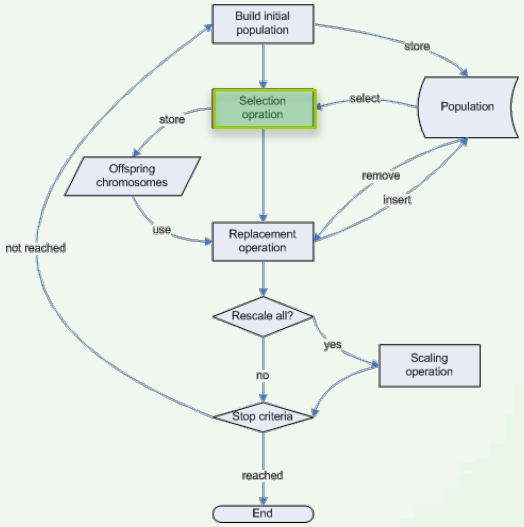
Evolution Process

- Need evaluation function
- Evaluate function and build “Roulette Wheel”

Roulette Wheel

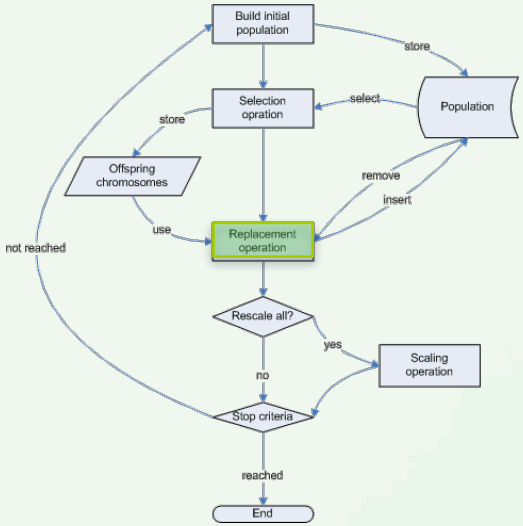


- Select next parents using Roulette Wheel method



Evolution Process

- New population is mix of children and previous generation parents
  - Chosen by roulette selection method



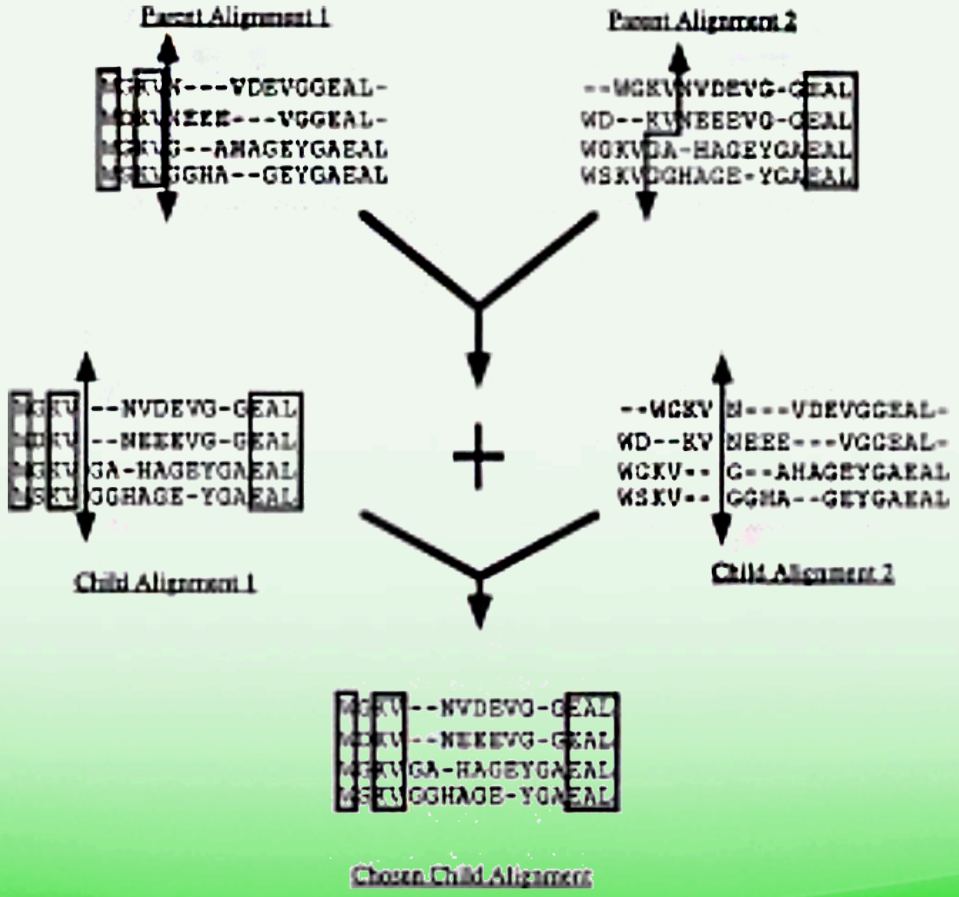
GAs in Sequence alignment

- Evaluation function

$$\text{ALIGNMENT COST(A)} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} W_{i,j}\, \text{COST}(A_i, A_j)$$

- Treat specific alignments as organisms in the population
- Recombine pieces of alignment to form new population

New Sequence Creation



Pseudocode

Initialisation	1. create G_0
Evaluation	2. evaluate the population of generation n (G_n)
	3. if the population is stabilised then END
	4. select the individuals to replace
	5. evaluate the expected offspring (EO)
Breeding	6. select the parent(s) from G_n
	7. select the operator
	8. generate the new child
	9. keep or discard the new child in G_{n+1}
	10. goto 6 until all the children have been successfully put into G_{n+1}
	11. n = n+1
	12. goto EVALUATION
End	13. end

Using Saga

- Search “SAGA multiple sequence alignment”
  - “tcoffee” webpage has a UNIX file
- Unzip/Install tar as ourlined in documentation
- Use

  saga pep_file my_file.pep PARALLEL_GA 2 PARALLEL_EXCHANGE 10

Input

>2gbp
ADTRIGVTIYKYDDNFMSVVRKAIEQDAKAAPDVQLLMNDSQNDQSKQND
QIDVLLAKGVKALAINLVDPAAAGTVIEKARGQNVPVVFFNKEPSRKALD
SYDKAYYVGTDSKESGIIQGDLIAKHWAANQGWDLNKDGQIQFVLLKGEP
GHPDAEARTTYVIKELNDKGIKTEQLQLDTAMWDTAQAKDKMDAWLSGPN
ANKIEVVIANNDAMAMGAVEALKAHNKSSIPVFGVDALPEALALVKSGAL
AGTVLNDANNQAKATFDLAKNLADGKGAADGTNWKIDNKVVRVPYVGVDK
DNLAEFSKK

>6abp
NLKLGFLVKQPEEPWFQTEWKFADKAGKDLGFEVIKIAVPDGEKTLNAID
SLAASGAKGFVICTPDPKLGSAIVAKARGYDMKVIAVDDQFVNAKGKPMD
TVPLVMLAATKIGERQGQELYKEMQKRGWDVKESAVMAITANELDTARRR
TTGSMDALKAAGFPEKQIYQVPTKSNDIPGAFDAANSMLVQHPEVKHWLI
VGMNDSTVLGGVRATEGQGFKAADIIGIGINGVDAVSELSKAQATGFYGS
LLPSPDVHGYKSSEMLYNWVAKDVEPPKFTEVTDVVLITRDNFKEELEKK
GLGGK

Results

alignment: binding.ref_aln contains 7 sequences, length=500

```
2gbp  ADTRIGVTIYK----YDDNFMSVVRKAIEQDAKAAPD---------VQLL
6abp  -NLKLGFLVKQ----PEEPWFQTEWKFADKAGKDLG----------FEVI
2liv  EDIKVAVVG--AMSGPVAQYGDQEFTGAEQAVADIN-AKGGIKGNKLQIA
2lbp  DDIKVAVVG--AMSGPIAQWGIMEFNGAEQAIKDIN-AKGGIKGDKLVGV
1gcg  ADTRIGVTIYK----YDDNFMSVVRKAIEKDGKSAPD---------VQLL
1glg  ADTRIGVTIYK----YDDNFMSVVRKAIEQDAKAAPD---------VQLL
1dbp  -KDTIALVVST----LNNPFFVSLKDGAQKEADKL-G---------YNLV

2gbp  MNDSQ-NDQSKQNDQIDVLLAKGVKALA--INLVDPAAAGTVIEKARGQN
6abp  KIAV--PDGEKTLNAIDSLAASGAKGFV--ICTPDPKLGSAIVAKARGYD
2liv  KYDD-ACDPKQAVAVANKVVNDGIKYVIGHLCSSSTQP---ASDIYEDEG
2lbp  EYDD-ACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQP---ASDIYEDEG
1gcg  MNDSQ-NDQSKQNDQIDVLLAKGVKALA--INLVDPAAAGTVIEKARGQN
1glg  MNDSQ-NDQSKQNDQIDVLLAKGVKALA--INLVDPAAAGTVIEKARGQN
1dbp  VLDSQ-NNPAKELANVQDLTVRGTKILL--INPTDSDAVDNAVKMANQAN

2gbp  VPVVFFNKEP----------SRKALDSYDKA-------YYVGTD-SKESG
6abp  MKVIAVDDQFVNAKGKPMDT----------V-------PLVMLA-ATKIG
2liv  ILMITPAATAPELT----------------ARGYQLILRT-TGLDSDQG
2lbp  ILMISPGATAPELT----------------QRGYQHIMRT-AGLDSSQG
1gcg  VPVVFFNKEP----------SRKALDSYDKA-------YYVGTD-SKESG
1glg  VPVVFFNKEP----------SRKALDSYDKA-------YYVGTD-SKESG
1dbp  IPVITLDRQA----------T--KG-EV--V-------SHIASD-NVLGG
```