Searching for similar proteins in a Database

BLAST

HMM

Threading

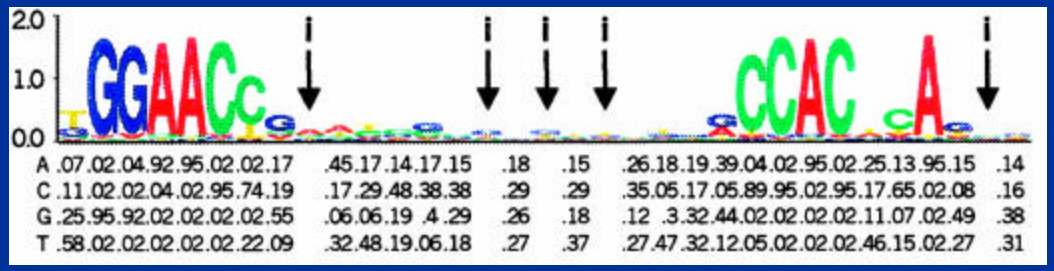
HMMER

Profile hidden Markov models for biological sequence analysis

http://hmmer.wustl.edu

```
40ptsH          TTTTGTGGCCTGCTTCAAACTT
41ptsH          TTTTATGATTTGGTTCAATTCT
42rhaS          AATTGTGAACATCATCACGTTC
43rot           TTTTGTGATCTGTTTAAATGTT
```

Alignment



Model

GTGTGATCAGAGTGATTGTGTCAGTGTGTAGCGCTCTGTT
TCGTGTGTTTGTGTTCATTTATTGTGTTGT GGCTTCTCATT
GCCCCTTTGGTTCTGTTCTTAAACCTTCATCTTCGCTTAGT
AAAGTTAGATTCCACCGA TCCGTTTCTGTTA AAGAAAAAG
TGATCAACAAACTTCAAGAAAATCTAAATGTGCAGTAATTT
GAAATTTATGCTTATTGTGT

Search for matches

The programs in HMMER:

- **hmmpfam**

  Search an HMM database for matches to a query sequence.

- **hmmbuild**

  build a model from a multiple sequence alignment

- **hmmsearch**

  Search a sequence database for matches to an HMM

HMM libraries:

PFAM: http://pfam.wustl.edu/
An HMM library based on the Swissprot 40 and SP-TrEMBL 18 protein sequence databases. 3882 protein families in current version.

SMART: http://smart.embl-heidelberg.de/
More than 500 extensively annotated domain families

The input and output:

MLYQLSKATTRIRLKRQKAVPQHRWLWSLAFLAAFTLKVSERANKNMAKTHNSGDVRCADLAI
SIPNNPGLDDGASYRLDYSPPFGYPEPNTTIASREIGDEIQFSRALPGTKYNFWLYYTNFTHHD
WLTWTVTITTAPDPPSNLSVQVRSGKNAIILWSPPTQGSYTAFKIKVLGLSEASSSYNRTFQVN
DNTFQHSVKELTPGATYQVQAYTIYDGKESVAYTSRNFTTKPNTPGKFIVWFRNETTLLVLWQ
PPYPAGIYTHYKVSIEPPDANDSVLYVEKEGEPPGPAQAAFKGLVPGRAYNISVQTMSEDEISL
PTTAQYRTVPLRPLNVTFDRDFITSNSFRVLWEAPKGISEFDKYQVSVATTRRQSTVPRSNEPV
AFFDFRDIAEPGKTFNVIVKTVSGKVTSWPATGDVTLRPLPVRNLRSINDDKTNTMIITWEADPA
STQDEYRIVYHELETFNGDTSTLTTDRTRFTLESLLPGRNYSL

	Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!!	fn3	139	221	1	84	58.1	1.2e-14	glocal	Fibronectin type III domain
!!	fn3	233	317	1	84	59.4	5.1e-15	glocal	Fibronectin type III domain
!!	fn3	328	410	1	84	36.3	4.4e-08	glocal	Fibronectin type III domain
!!	fn3	421	501	1	84	58.4	9.8e-15	glocal	Fibronectin type III domain
!!	fn3	512	591	1	84	27.0	3e-05	glocal	Fibronectin type III domain
!!	fn3	599	677	1	84	78.9	6.9e-21	glocal	Fibronectin type III domain
!!	fn3	689	778	1	84	40.8	2e-09	glocal	Fibronectin type III domain
!!	fn3	789	869	1	84	14.8	0.0063	glocal	Fibronectin type III domain
!!	fn3	880	955	1	84	67.6	1.7e-17	glocal	Fibronectin type III domain
!!	fn3	974	1060	1	84	58.4	1e-14	glocal	Fibronectin type III domain
!!	Y_phosphatase	1312	1542	1	274	393.6	1.3e-115	glocal	Protein-tyrosine phosphatase

Program 1: hmmpfam

Search an HMM database for matches to a query sequence.

Query sequence:
C:\cbsu\module1\hmmer_projects\exe1\unknown_proteins

Database:
pfam

Exercise 1: Identifying domains in an unknown protein

1. Check the files in the directory

   C:\cbsu\module1\hmmer_projects\exe1

   - sequence file: unknown_proteins
   - database file: pfam_test
   - program file: hmmpfam.exe, parse_pfam.pl

2. Run the program: (for help: hmmpfam –h)

   hmmpfam -E 1e-10 -A 10 pfam_test unknown_proteins > pfamresult.txt

3. Parse the result into a spreadsheet

   parse_pfam.pl pfamresult.txt pfamresult.xls

Evaluating the significance of a hit:

1. E-value: <= 0.1
(10% chance that you would've seen a hit this good in a search of random sequences)

2. Raw score >= GA (the scores used as cutoffs in constructing Pfam)

3. Raw score > $\log_2$(number of seqs in the database) **(20 for the nr)**

Parallel HmmPfam at CBSU

contact: cbsu@tc.cornell.edu

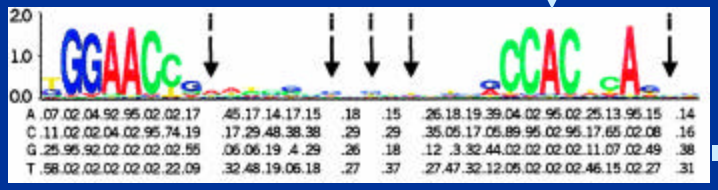
Program 2, 3. hmmbuild & hmmsearch

Build a model and search the sequence database for motifs that fit the model.

```
40ptsH          TTTTGTGGCCTGCTTCAAACTT
41ptsH          TTTTATGATTTGGTTCAATTCT
42rhaS          AATTGTGAACATCATCACGTTC
43rot           TTTTGTGATCTGTTTAAATGTT
```

Sequence alignment

hmmbuild



Model

hmmsearch

More sequence motifs that fit this model

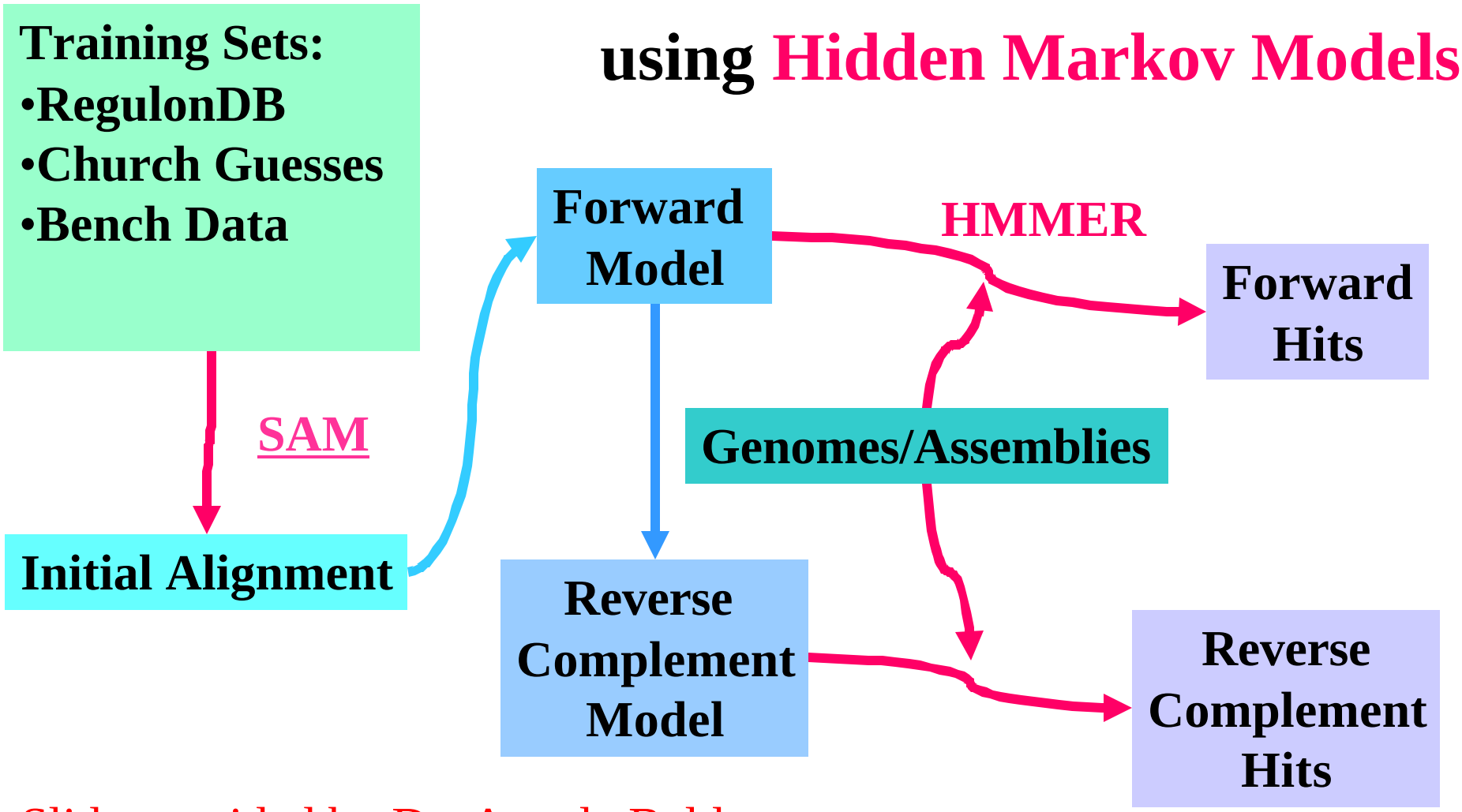
Exercise 2: Identifying all putative genes that are regulated by crp (Cyclic AMP receptor)

Available resources:

1. E coli genome sequence

2. a list of crp binding sites determined by DNA footprinting.

Automated Process: Finding Known Regulators in Genome Sequences using Hidden Markov Models



- Slide provided by Dr. Angela Baldo

Exercise 2: Building models

1. Check the files in the directory.

   C:\cbsu\module1\hmmer_projects\exe2

   - sequence alignment: crp-Church.aln
   - database file: ecoli_k12
   - program file: hmmbuild.exe, hmmcalibrate.exe, hmmsearch.exe

2. Build the model: (for help: hmmbuild –h)

   hmmbuild crpmodel crp-Church.aln

3. Calibrate the model:

   hmmcalibrate crpmodel

4. Search the genome: (for help: hmmsearch –h)

   hmmsearch crpmodel ecoli_k12 > searchresult.txt