EVOLUTIONARY MODEL DESCRIPTION LANGUAGE.

SOME NECESSARY FEATURES

SERGEI L KOSAKOVSKY POND (SPOND@UCSD.EDU)

EMDL?

A USEFUL EVOLUTIONARY MODEL DESCRIPTION LANGUAGE SHOULD BE ABLE TO ENCAPSULATE A COMPLETE, SELF-CONTAINED DESCRIPTION OF A PROBABILISTIC MODEL WHICH, WHEN GIVEN A NUMBER OF MODEL-SPECIFIED INPUTS, SUCH AS SEQUENCE ALIGNMENTS, PHYLOGENETIC TREES, STRUCTURAL OR GENE PARTITIONS CAN COMPUTE A FIT OF THE MODEL TO THE DATA (E.G. MAXIMUM LIKELIHOOD) AND GENERATE INTERPRETABLE VALUES AND STATISTICAL DESCRIPTIONS OF ESTIMATED MODEL PARAMETERS (E.G. BRANCH LENGTHS, DN/DS).

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

MAP EACH INPUT CHARACTER TO A STATE DISTRIBUTION IN THE MODEL SPACE, E.G.

1). NUCLEOTIDE/PROTEIN

'A' → (1,0,0,0) 'R' → (1,0,1,0) 'F' → Ø

2). CODON/DINUCLEOTIDE

'AAG' → (0,0,1,...), 'TAG' → Ø

3). MICROSATTELITE

5 → (0,0,0,0,0,1,...)

4). 3-STATE HIDDEN STATE 'COVARION' MODEL FOR NUCLEOTIDES

'A' → (1,0,0,0,1,0,0,0,1,0,0,0)

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

WHAT NEEDS TO BE FORMALIZED

1. STATE SPACES FOR PUBLISHED MODELS
2. STANDARD MAPPING FUNCTIONS FROM SEQUENCE DATA TYPES TO MODEL STATE SPACES
3. EXCEPTION HANDLING CONVENTIONS (E.G. WHAT TO DO IF A STOP CODON IS INPUT TO A CODON MODEL, HOW TO HANDLE INDELS)
4. GENERAL MAPPING MECHANISM (E.G. 'STRING' TO 'INTEGER')

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

AN EVOLUTIONARY MODEL CONTAINS PARAMETERS

1. TIME PARAMETERS
2. RATE PARAMETERS
3. FREQUENCY PARAMETERS
4. 'OTHER' PARAMETERS

PARAMETERS CAN BE

1. FIXED
2. ESTIMATED
3. CONSTRAINED & ESTIMATED

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

WHAT NEEDS TO BE FORMALIZED

1. PARAMETER IDENTIFIERS/ NAMESPACES AND REFERENCE MECHANISM
2. PARAMETER SCOPE
   - 2A BRANCH
   - 2B TREE
   - 2C PARTITION
   - 2D GLOBAL
3. ESTIMATION POLICY
   - 3A FIXED
   - 3B ESTIMATED
   - 3C STOCHASTIC/INTEGRATED
   - 3D ...
4. PARAMETER CONSTRAINTS, E.G. FREQUENCIES SUM TO ONE

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

DEFINE THE PROBABILITY OF SUBSTITUTING ONE CHARACTER WITH ANOTHER GIVEN THE VALUES OF ALL MODEL PARAMETERS (CONSIDER ONLY TIME-HOMOGENEOUS MODELS FOR NOW)

1. DEFINE THE RATE/TRANSITION MATRIX
2. DESCRIBE PROCESS PROPERTIES (E.G. TIME-REVERSIBILITY) THAT CAN AFFECT ESTIMATION AND MAY NOT BE EASILY INFERRED AT RUN TIME

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

MODEL

ALPHABET

MODEL PARAMETER DESCRIPTION

CHARACTER SUBSTITUTION PROCESS

WHAT NEEDS TO BE FORMALIZED

1. HOW TO COMPACTLY DEFINE THE RATE/TRANSITION MATRIX ON THE CHARACTER ALPHABET
   - 1A. CANONICAL MARKOV PROCESS FORMS
   - 1B. ENUMERATION FOR SMALL MODELS (NUCLEOTIDES 4X4)
   - 1C. 'CASES' FOR MORE COMPLEX MODELS (CODON)
2. HOW TO REFERENCE MODEL PARAMETERS IN THE RATE MATRIX DEFINITION
3. WHAT ARE SOME IMPORTANT PROCESS PROPERTIES THAT NEED TO BE 'TAGGED'.

BASIC COMPONENTS OF AN EVOLUTIONARY MODEL

NEXT

- [ ] MODELS BY THEMSELVES ARE NOT IMMEDIATELY USEFUL
- [ ] NEED TO BE COMBINED IN AN ANALYSIS, E.G. ASSIGNED TO TREES/PARTITIONS, CONSTRAINTS DEFINED, ESTIMATION PROCEDURES DESCRIBED, HIGHER ORDER PARAMETERS DEFINED (E.G. MODEL MIXTURE PROPORTIONS, DISTRIBUTIONS)
- [ ] RESULTS MUST BE PREPARED AND REPORTED

USEFUL TEST-CASES

- [ ] 4X4 NUCLEOTIDE MODEL, WITH OR WITHOUT RATE VARIATION, E.G. GTR+G+I
- [ ] SITE CODON MODELS OF YANG ET AL GENETICS (2000) 155(1): 431-49.
- [ ] BRANCH CODON MODELS OF NIELSEN AND YANG MBE (2002) 19:908-917
- [ ] DEFINE STEM-RNA MODELS OF SAVILL ET AL GENETICS (2001) 157(1):399-411
- [ ] MIXTURE AMINO-ACID MODELS, E.G. THE FITNESS MODELS OF DIMMIC ET AL
- [ ] COVARION CODON MODELS GUINDON ET AL PNAS (2004) 101 (35):12957-62
- [ ] CODON MIXTURE MODELS, SUCH AS WHELAN & GOLDMAN GENETICS (2004) 167: 2027-2043
- [ ] HIDDEN MARKOV RATE MODELS OF FELSENSTEIN AND CHURCHILL MBE (1996) 13: 93-104