Genes and gene finding

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

Central dogma

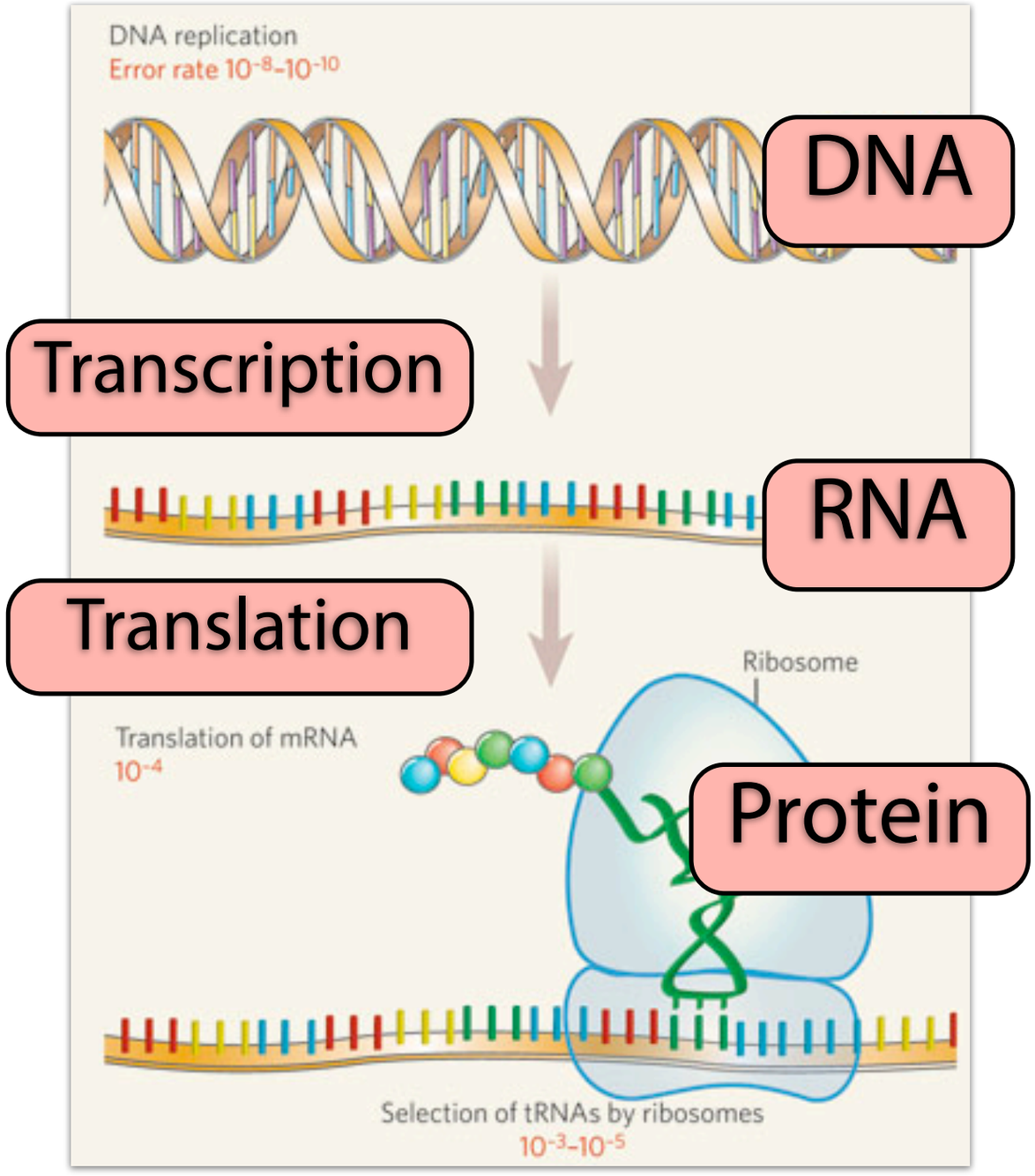
Short version:

DNA -> RNA -> Protein

Long version:

DNA molecules contain information about how to create proteins; this information is *transcribed* into RNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.



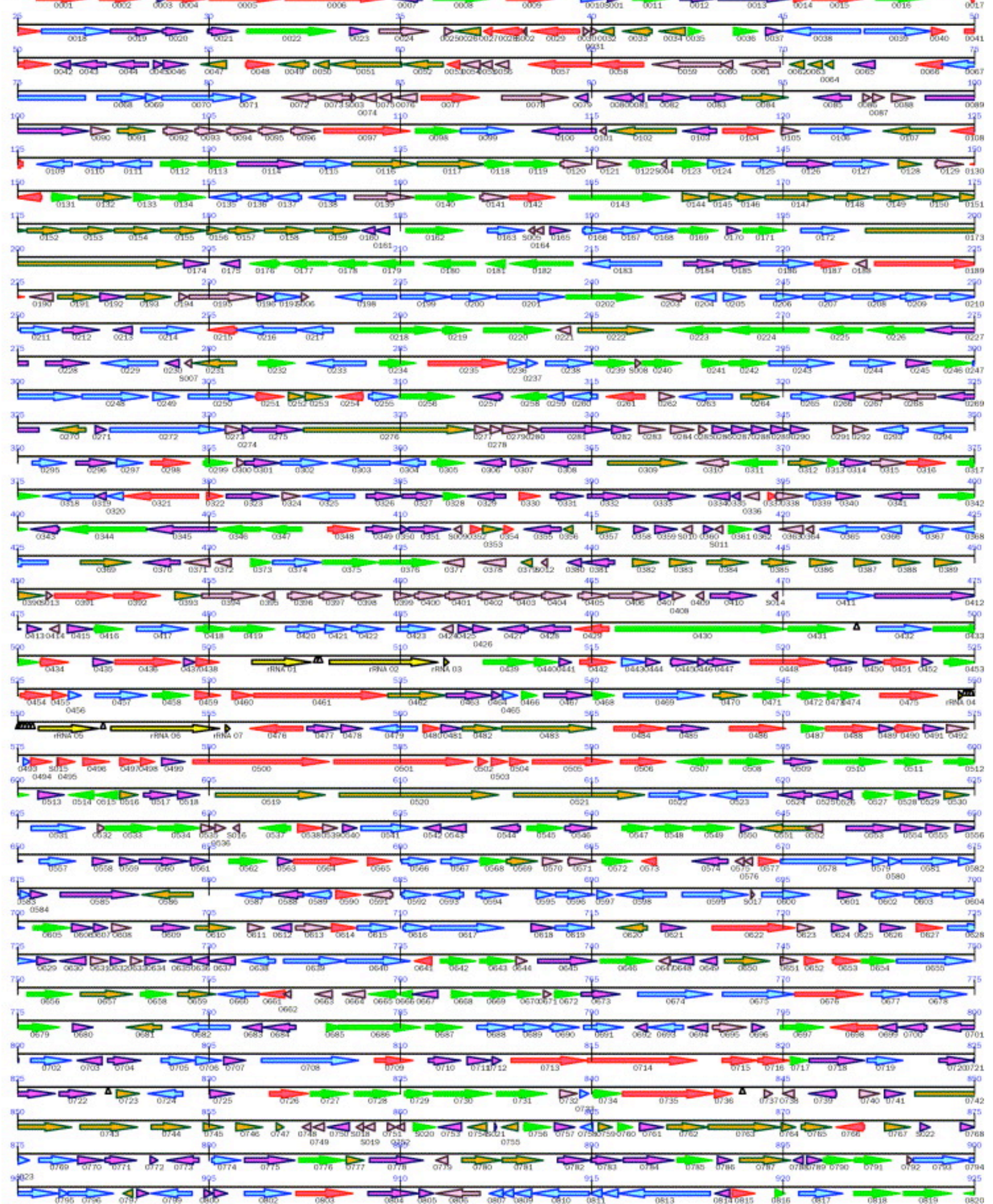
Picture from: Roy H, Ibba M. Molecular biology: sticky end in protein synthesis. Nature. 2006 Sep 7;443(7107):41-2.

Bacterial genes

Genes (colored arrows) packed tightly into the *Staphylococcus aureus* genome

In bacteria, one gene corresponds to one continuous interval on the genome

We have good methods for predicting where these genes are



Bacterial gene finding

Table 3. Glimmer3 prediction accuracy with `long-orfs` training

Genome			Glimmer3 Predictions					versus Glimmer2.13		
Organism	GC%	# Genes	3′ Matches		5′ & 3′ Matches		Extra	3′ Match	5′ & 3′	Extra
A.fulgidus	49	1165	1161	99.7%	873	74.9%	1332	−2	−34	−64
B.anthracis	35	3132	3125	99.8%	2751	87.8%	2419	−1	+752	−144
B.subtilis	44	1576	1562	99.1%	1391	88.3%	3020	+3	+421	−724
C.tepidum	57	1292	1289	99.8%	934	72.3%	835	+3	+26	−400
C.perfringens	29	1504	1501	99.8%	1383	92.0%	1192	−1	+267	−20
E.coli	51	3603	3534	98.1%	3112	86.4%	1002	+11	+784	−843
G.sulfurreducens	61	2351	2337	99.4%	1933	82.2%	1165	+7	+575	−734
H.pylori	39	915	910	99.5%	795	86.9%	788	+2	+57	−103
P.fluorescens	63	4535	4510	99.4%	3598	79.3%	1953	+35	+895	−2359
R.solanacearum	67	2512	2485	98.9%	2028	80.7%	1183	+341	+1044	−2184
S.epidermidis	32	1650	1646	99.8%	1514	91.8%	791	+8	+358	−32
T.pallidum	53	575	567	98.6%	391	68.0%	567	−2	+50	−281
U.parvum	26	327	324	99.1%	295	90.2%	297	−1	+21	−11
Averages:				99.3%		83.1%		+31	+401	−608

Genomes and columns are as in the preceding table. Glimmer3 was run by using the output of its `long-orfs` program to train an IMM. The output of an initial run of Glimmer3 was used to set start codon frequencies and to find a ribosome-binding-site motif. A second run of Glimmer3 using those values generated the above predictions. Glimmer2 was trained on the output of its version of the `long-orfs` program.

Approaches can identify exact bacterial genes with as high as 92% accuracy; can identity gene ends with ≥98% accuracy

Delcher, Arthur L., et al. "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics* 23.6 (2007): 673-679.

Eukaryotic genes

Eukaryotic genes are more complex than prokaryotic (bacterial) genes for several reasons, as we'll see

Likewise, *finding* eukaryotic genes computationally is harder

A human gene

chr11:5246500-5248500 (reverse strand):

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA
TACCTCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTT
TCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTT
GAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTA
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA
TCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

A human gene

chr11:5246500-5248500 (reverse strand):

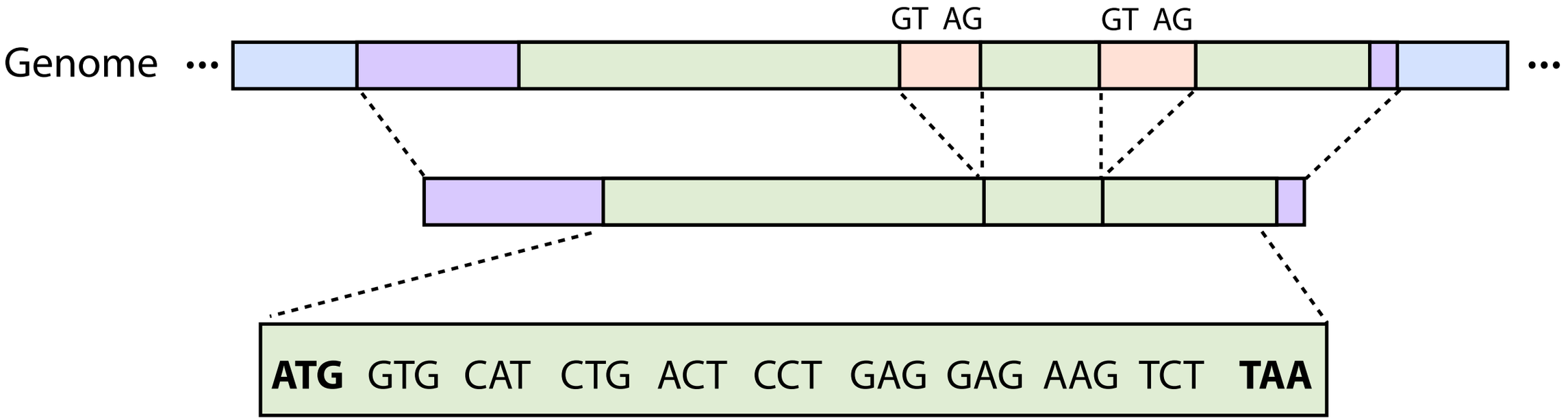
ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA
TACCTCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTT
TCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTT
GAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTA
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA
TCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

Homo sapiens hemoglobin, beta (HBB)

Transcription

Can an HMM help us find eukaryotic genes and their constituent parts?

Parts: non-genes, exons (both coding and non-coding portions), introns



Sequence signals: acceptors, donors, branch sites, pyrimidine-rich sites, nucleotide compositions

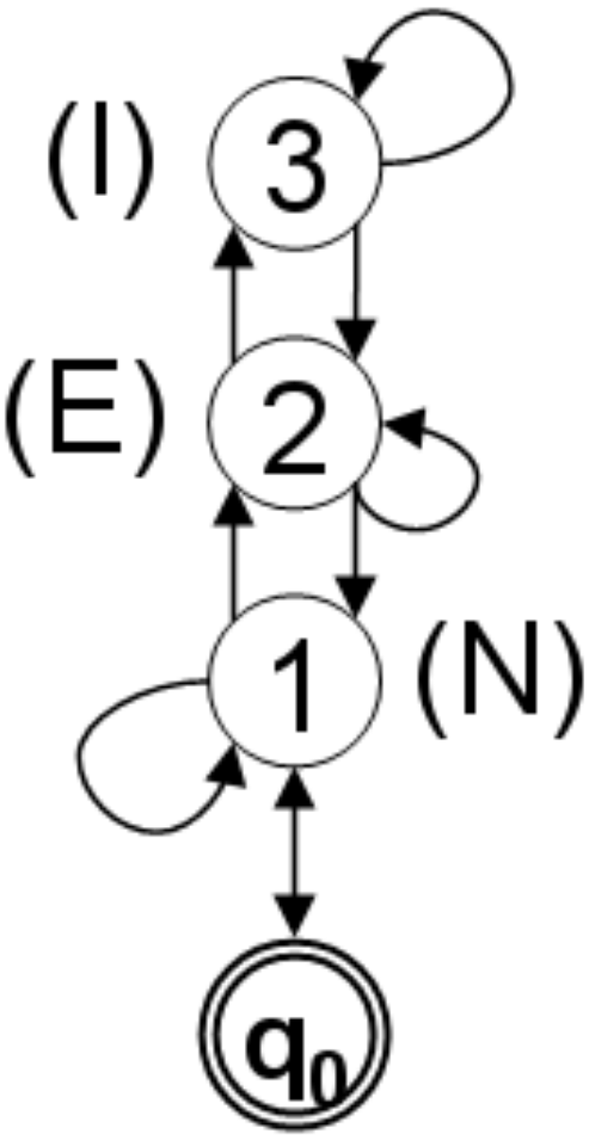
Eukaryotic gene finding

Attempt 1:

Emissions are nucleotides

I = intron
E = exon
N = intergenic
(between genes)

q0 is a *start state*; guarantees we start in the N (intergenic) state



Model captures:

Genes, exons and introns

Does not capture:

Start/stop codons, acceptors/donors, codons

Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Attempt 2:

Three exon states
additionally capture *codons*

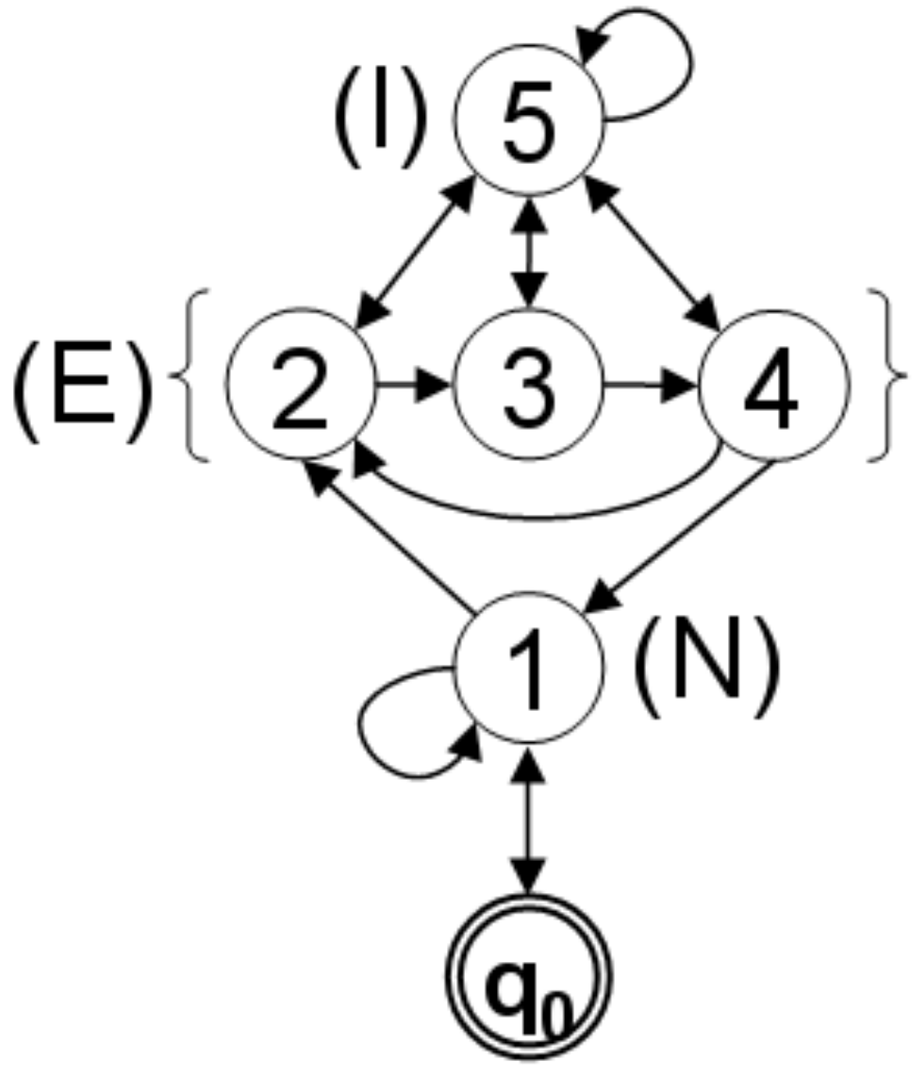


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Attempt 3:

States 2-4 capture start codon

13, 14 capture donor

16, 17 capture acceptor

8-12 capture stop codons

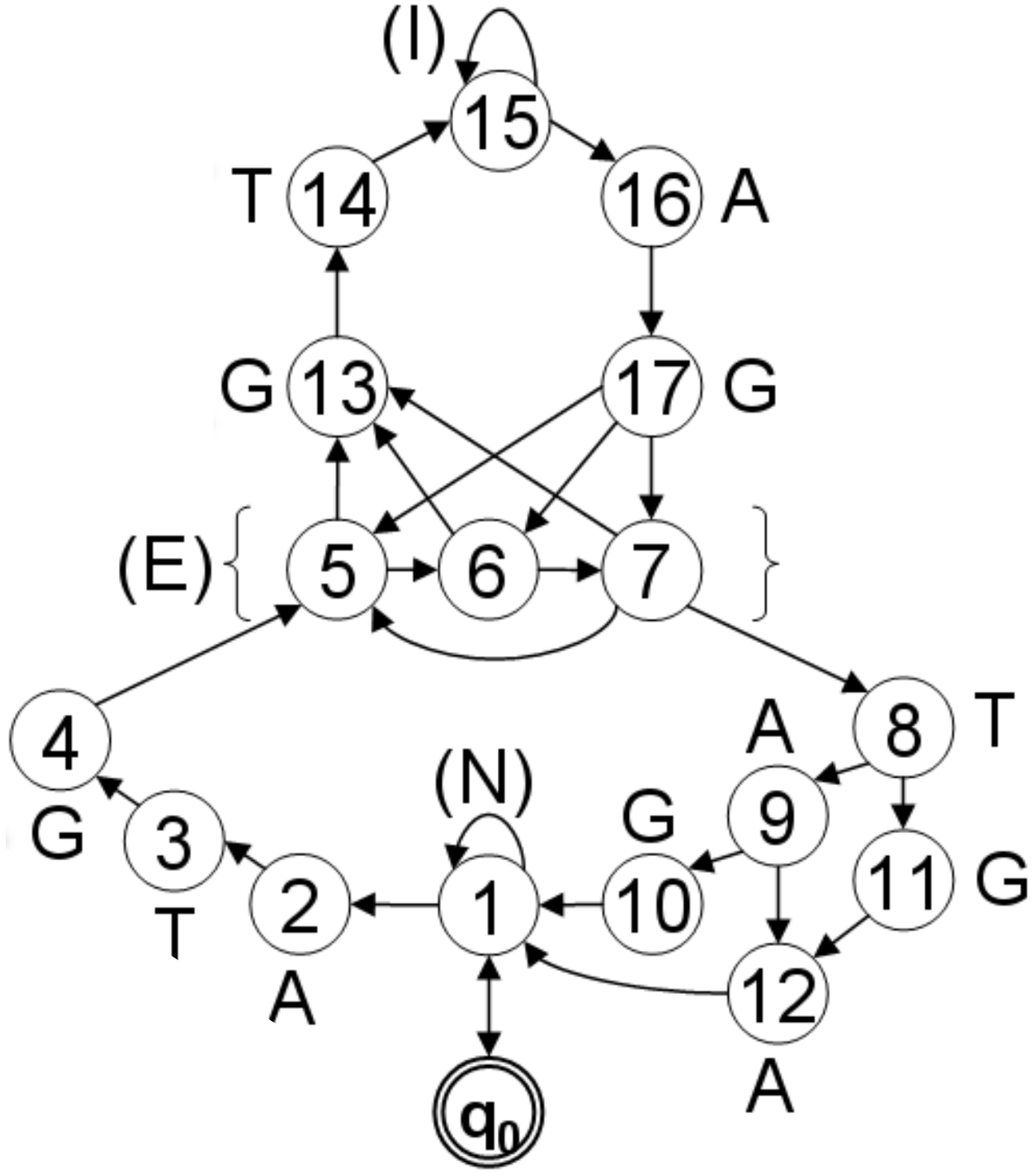


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Attempt 4:

Additional copies of the accepter/intron/donor loop allow us to pick up where we left off in reading frame

Recall transition probability matrix has $|Q|^2$ elements

27 states → 729 transition probs

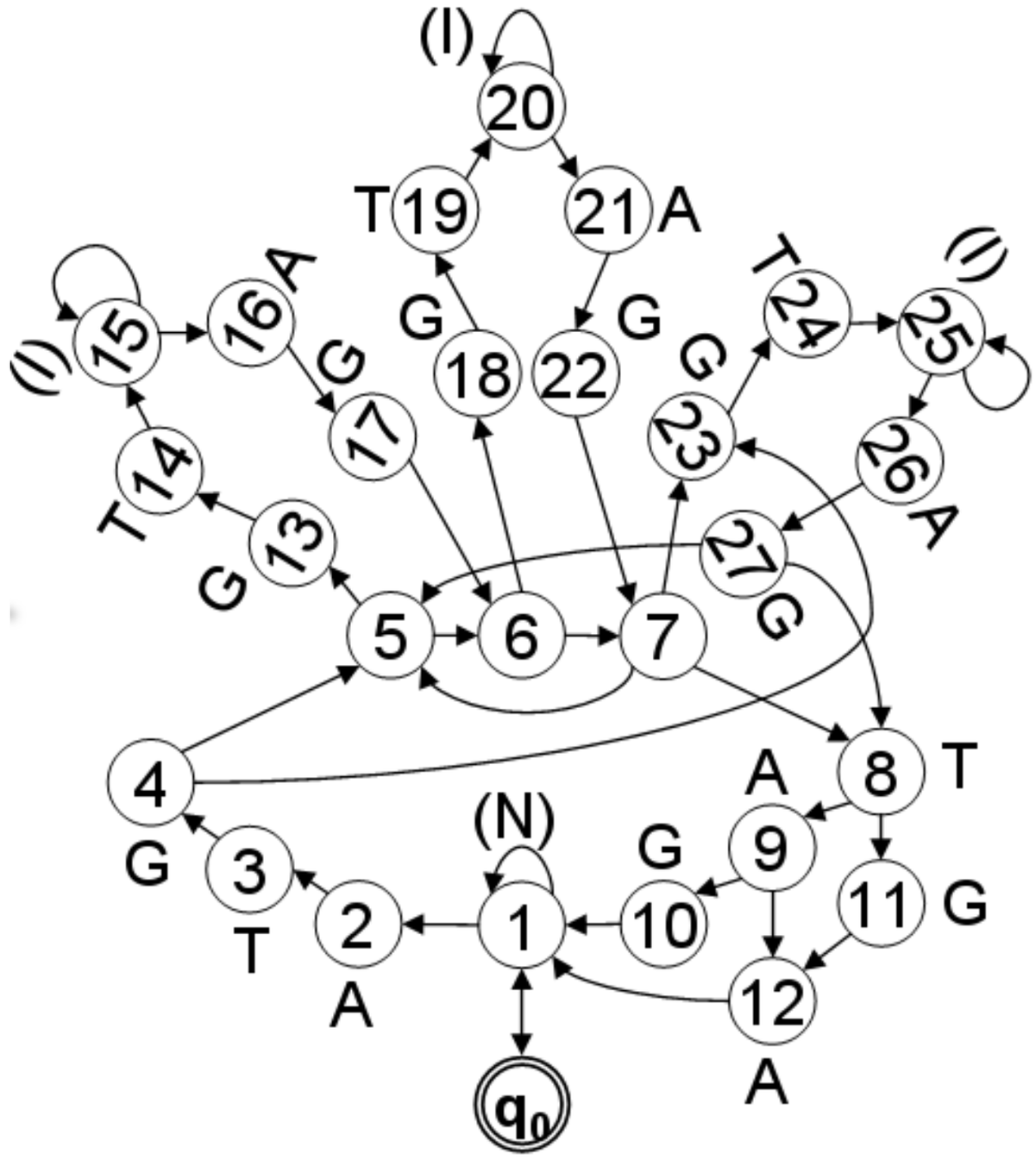


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.