Combinatorics of Genome Rearrangements

Guillaume Fertin, Anthony Labarre, Irena Rusu, Éric Tannier and Stéphane Vialette

The MIT Press
Cambridge, Massachusetts
London, England

© 2009 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Times New Roman and Syntax on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Combinatorics of genome rearrangements / Guillaume Fertin . . . [et al.].
p. cm. — (Computational molecular biology)
Includes bibliographical references and index.
ISBN 978-0-262-06282-4 (hardcover : alk. paper) 1. Translocation (Genetics)—Mathematical models.
2. Translocation (Genetics)—Data processing. 3. Combinatorial analysis. 4. Genomics—Mathematics.
I. Fertin, Guillaume, 1972– II. Series.
[DNLM: 1. Gene Rearrangement. 2. Genome. 3. Models, Genetic. QU 470 C731 2009]
QH462.T7C66 2009
572.8′77—dc22 2008042152

10 9 8 7 6 5 4 3 2 1

Index

2-break rearrangement, 73
2-gene, 203
k-break rearrangement, 172
3-DIMENSIONAL MATCHING, 252
3-PARTITION, 252
3-SATISFIABILITY, 253

ALTERNATING CYCLE DECOMPOSITION, 253
BIN-PACKING, 253
BREAKPOINT GRAPH DECOMPOSITION, 253
BREAKPOINT PHYLOGENY PROBLEM, 210
CLIQUE, 253
EULERIAN CYCLE DECOMPOSITION, 253
EXACT COVER BY 3-SETS, 253
FULL STEINER TREE PROBLEM, 208
HAMILTONIAN CIRCUIT, 254
HAMILTONIAN PATH, 254
INDEPENDENT SET, 254
LARGE PARSIMONY PROBLEM, 208
LARGEST BALANCED INDEPENDENT SET, 254
LONGEST PATH PROBLEM, 237
MAXIMUM PARSIMONY ON BINARY ENCODINGS, 218
MINIMUM COMMON STRING PARTITION, 133
REVERSAL PHYLOGENY PROBLEM, 211
REVERSED MINIMUM COMMON STRING PARTITION, 137
SATISFIABILITY, 254
SEQUENCE-BASED MAXIMUM PARSIMONY PROBLEM, 218
SET PACKING, 254
SHORTEST PATH PROBLEM, 237
SIGNED MINIMUM COMMON STRING PARTITION, 136
SMALL PARSIMONY PROBLEM, 212
STEINER TREE PROBLEM, 208
TRANSPOSITION PHYLOGENY PROBLEM, 211
TRAVELING SALESMAN, 255
VERTEX COVER, 255

Adjacency, 20, 162
  belongs, 173
  in a genome, 162
  matrix, 239
  reverse, 20
  telomeric, 162
  vertices, 235
Alphabet, 92
Alternating group, 15
Arc, 240
  black, 28
  desire, 28
  gray, 28
  head, 240
  reality, 28
  tail, 240
Ascent, 27

Block, 93
  edit, 93
  family, 124
  move, 128
  uncopy, 128
  uniform, 141
Block interchange
  permutations, 49
  strings, 153
Breakable, 172
Breakpoint, 20
  binary string, 139
  distance, 21
  external, 162, 163
  internal, 162, 163
  phylogeny, 210
  prefix transposition, 37
  strong, 20
Breakpoint diagram, 35
Breakpoint graph
  contracted, 201
  of a genome, 167
  of permutation, 41
  signed permutation, 65

Cayley graph, 17
Center, 206
Centromere, 168

Character encodings
  MPBE, 218
  MPME, 218
  SB encoding, 218
Character moving, 151
Character swaps, 155
Chromosome, 1, 162
  circular, 162
  legal, 169
  linear, 162
Circular lower bound, 214
Clan, 38
Clique, 235
Collection, block edit, 123
Complement of a graph, 235
Component
  genomes, 164
    chain, 173
    tree, 173
  minimal, 164
  oriented, 173
  of an overlap graph, 86
  permutation, 66
    chain, 66
    tree, 66
Conjugacy class, 14
Connected
  component, 237
  graph, 237
Cover
  cost, 174
  of a poset, 76
  of a PQ-tree, 173
$\mathcal{CS}$-Factor, 113
Cut-and-paste, 55
Cycle
  even, 29
  graph, elementary, 237
  in a graph, 236
    even, 237
    length, 236
    odd, 237
  Hamiltonian, 237
  interleaving, 83
  length, 30
  odd, 29
  of a permutation, 14

DCJ, 170
  cut, 170
  join, 170
Degree
  of a graph
    maximum, 236
    minimum, 236
  of a vertex, 236
Deletion
  block, 94
    exclusive, 102
  of a nucleotide, 2
  of a segment, 3
Dense instance, 140
Descent, 27
Diameter, 6
  of a graph, 237
  of a metric space, 6
  syntenic, 185
Digraph
  connected
    strongly, 241
    weakly, 241
Directed acyclic graph, 241
  topological sort, 241
  transitive closure, 242
Directed graph, 240
  acyclic, 241
  cycle, 241
Disjoint cycle decomposition, 14
Distance
  k-break, 172
  between two genomes, 6
  between vertices, 237
  block covering, 124
  block edit, 123
    exclusive, with reversals, 102, 103
  breakpoint, 163
    poset, 80
  Cayley, 50
  conserved interval, 64, 115
  DCJ, 170
    permutation, 73
  double, 201
  Hamming, 123
  left-invariant, 18
  Levenshtein, 123
  monotonic, 104
  prefix exchange, 52
  prefix transposition, 37
  reversal, 40
    perfect, 69
    poset, 79
    prefix, 70
  RT, 172
  string edit, 123
  syntenic, 182
    linear, 186
  TDRL, 59
  translocation, 165, 166
  transposition, 26
  Ulam's, 31
DNA, 1
  complementarity, 1
  segment, 1
  strand, 1
Double cut-and-join, 73, 170
Dual genome representation, 183

Duplicate, 92
Duplication, 3, 91
  block, 94
  tandem, 3

Edge, 235
  desire, 65
  extremities, 235
  interleaving, 83
  reality, 65
Edit
  block edit, 93
  character edit, 123
Element of a chromosome, 169
Elimination graph, 85
Eulerian
  cycle, 237
  path, 237
Evolution, 2
Extremity
  edge, 235
  gene, 161
  interval, 163, 164
  path, 236

Fission, 4, 159, 165, 182
Forest, 238
Function
  goal, 246
  objective, 246
Fusion, 4, 159, 165, 182

Gene, 1, 13, 92
  duplicated, 200
  extremity, 161
  family, 91
  head, 65, 161
  signed, 161
  tail, 161
  unsigned, 161
Generator, 16
Genome, 1
  compact representation, 182
  doubled, 200, 203
  duplicated, 200, 203
  halving, 201
  linear, 162
  median, 193
  set system, 203
  unsigned, 161
Genome rearrangement problem, 6
Graph
  Γ-, 30
  k-regular, 236
  acyclic, 237
  bipartite, 238
    balanced, 238
    complete, 238
  breakpoint, 65
  Cayley, 17
  complement of a, 235
  complete, 235
  connected, 237
  cycle, 28
  diameter, 237
  digraph, 240
  directed, 240
  disconnected, 237
  interleaving, 83
  loop, 235
  order of a, 235
  overlap, 84
  path, 236
  of a permutation, 14
  simple, 235
  size of a, 235
  synteny, 184, 203
  undirected, 235
Group
  alternating, 15
  hyperoctahedral, 15
  symmetric, 14
Grouping, 138
Guided halving, 202

Hamiltonian
  cycle, 237
  path, 237
Hasse diagram, 77
Homology, 4, 91

Identity genome, 166
In-degree, 240
  maximum, 241
  minimum, 240
Incidence, 235
Independent
  edges, 235
  vertices, 235
In-neighborhood, 240
Insertion
  block, 94
    exclusive, 102
  of a nucleotide, 3
Interleaving
  cycles, 83
  edges, 83
  graph, 83
Interval
  2-interval, 110
  active, 44
  break, 69
  common, 21, 164
    of two strings, 113
  conserved, 64, 113, 164
    of a set of permutations, 115
  extremities of an, 21
  genome, 163

Interval (cont.)
 linear, 23
 (maximal) location of, 112
 nested, 69
 passive, 44
 of a permutation, 21
 prime, 23
 separable, 70
 of a string, 113
Inversion, 3

Kendall's tau, 51

Leaf, 238
Left code
 of an element, 31
 of a permutation, 31
Left-invariance, 18
Legal
 chromosome, 169
 rearrangement, 169
Length of a cycle in a graph, 236
Linear
 rearrangement, 172
 set, 23
Linear extension
 of a permutation, 20
 of a poset, 76
 of a string, 136
Linearly ordered set, *see* Totally ordered set
Local complementation, 85
 game, 85
Longest common prefix, 93
Longest common subsequence, 93
Longest common substring, 93
 up to a reversal, 93
Longest common suffix, 93

Matching, 238
 t-dimensional, 115
 exemplar, 99
 full, 99
 intermediate, 99
 maximal, 238
 maximum, 238
 perfect, 238
Matrix adjacency, 239
Measure
 exemplar, 99
 full, 99
 intermediate, 99
Median
 genome, 193
 problem, 193
 score, 193
Metric, 6
 space, 6

Model
 exemplar, 98
 full, 98
 intermediate, 98
MPBE, 218
Mutation point, 2

Neighborhood, 235
Node, vertex of a graph, 235
Noncomparability, 76
NP, 243
Nucleotide, 1
Number
 MAD, 118
 SAD, 118

Oriented
 component of a genome, 173
 component of an overlap graph, 86
 component of a permutation, 66
 cycle, 84
 edge, 84
 element, 85
 point, 66
 reversal, 68
 vertex, 84, 85
Orthology, 91
Out-degree, 240
 maximum, 241
 minimum, 240
Out-neighborhood, 240
Overlap
 graph, 84
 intervals, 22

P, 243
Paralogy, 91
Parameterized complexity, 250
Partially ordered set, 76
 antichain, 76
 chain, 76
 down-set, 77
 height, 76
 linear extension, 76
 maximal element, 77
 minimal element, 77
 up-set, 77
 weak order, 76
 width, 76
Partial order, 75
Partition
 common, 134
 common possibly reversed, 137
Path
 directed, 241
 in a graph, 236
 Hamiltonian, 237

Perfect
 matching, 238
 scenario, 69
Permutation, 13
 2-permutation, 32
 3-permutation, 32
 γ-permutation, 34
 k-permutation, 83
 circular, 19
 even, 15
 extended, 19
 genomic circular, 19
 Gollan, 40
 identity, 14
 inverse, 14
 linear, 13
 linearization, 19
 odd, 15
 pattern, 57
 reduced, 27
 reversed, 33
 signed, 15
 simple, 32
 spin of a, 43
 stack-sortable, 56
 subsequence, 31
  increasing, 31
 toric, 32
Phylogeny, 191
 breakpoint, 210
 reversal, 211
Plateau, 31
Point
 belonging to a component, 66
 in a permutation, 20
Polynomial time
 reducibility, 244
 reduction, 244
Poset. *See* Partially ordered set
PQ-tree
 components, 66
 intervals, 23
 P-node, 23, 66
 Q-node, 23, 66
Prefix transposition, 152
Prime interval, 23
Problem
 decision, 243
 optimization, 246
Pruning
 $\mathcal{M}$-pruning of two sets of permutations, 116
 t-dimensional, 116
 exemplar, 99
 full, 99
 intermediate, 99

Random loss, 58
Replacement block, 94
Replication, 2
 origin, 70
 terminus, 70
Retrotransposition, 3
Reversal, 3, 40
 k-reversal, 54, 71
 length, 52
 oriented, 68
 phylogeny, 211
 prefix, 47
 safe, 67
 signed, 64
 symmetric around a point, 70
Revrev, 73
Right code
 of an element, 31
 of a permutation, 31

Segment
 of DNA, 1
 of a permutation, 21
Sequence, 1
 alignment, 2
 sorting, 18
Similarity
 adjacency, 109
 common intervals, 111
 conserved intervals, 114
Singleton, 102
 strip, 43
Sink, 240
Size, genome, 181
Source, 240
Speciation, 91
Spin, canonical, 46
Stack, 56
 full-pop, 56
 pop, 56
 push, 56
 sorting, 56
 top, 56
Steiner
 set, 208
 tree, 208
  full, 208
  terminal, 208
Stirling number of the first kind, 51
String, 92
 h-span, 104
 k-ary, 92
 balanced, 94
 binary, 92
 circular, 93
 pegged, 102
 prefix, 92
 related, 94
 signed, 92
 subsequence, 93

String (cont.)
  substring, 92
  suffix, 93
  ternary, 92
Strip, 27
  move, 56
  strong, 43
    decreasing, 43
    increasing, 43
    long, 43
Strong interval, 22
Subgraph, 236
  induced, 236
  spanning, 236
Substitution of a nucleotide, 2
Synteny, 182
  exact, 188
  graph, 184
  linear, 186
    nested, 187
  uncovering, 188

Telomere, 1, 162
Telomeric
  adjacency, 162
  marker, 161
Tetraploidy, 202
Tight
  directed graph, 87
  graph, 85
  matrix, 85
Topological sorting, 77
Toric
  equivalence, 32
  permutation, 32
Totally ordered set, 76
Transfer
  horizontal, 4
  lateral, 4
Translocation, 164
  reciprocal, 4, 159, 164, 182
Transposition, 25
  k-transposition, 34, 37
  (p, q)-transposition, 55
  prefix, 37
Transreversal, 72
Tree, 238
  ordered, 238
  root of a, 238
  rooted, 238
  Steiner, 208
  strong interval, 22
Tree topology, 212

Vertex, 235
  child, 238
  class, 238
  elimination, 85
  extremity, 236
  isolated, 235
  matched, 238
  parent, 238
  reachable, 241
  sink, 240
  source, 240
  well-separated set, 204