

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABLES	iv
GLOSSARY	v
LIST OF ABBREVIATIONS	vii
LIST OF VARIABLES	viii
PREFACE	x
INTRODUCTION	1
CHAPTER 1: RANDOM SUBCLONING	5
1.1 Mapping And Sequencing Large Genomes	6
1.2 Sequence Walking	8
1.3 Overview Of Random Subcloning	12
1.4 Mathematical Model - Basic Formulation	18
1.5 Target Coverage (1)	21
1.6 Mathematical Model - The Beta Distribution	25
1.7 Gaps And Islands	29
1.8 The Number Of Clones In An Island	31
1.9 Island Length	35
1.10 Target Coverage (2)	37
1.11 Comparison With The Lander-Waterman Equations	40
1.12 Island Co-Dependency	44
1.13 Circular Targets	46
1.14 Simulations And Data	51
1.15 Examples	53
1.16 Random Closing Remarks	60
CHAPTER 2: PAIRWISE END SEQUENCING	64
2.1 The Double-Barrel Shotgun	64
2.2 Formulation	66
2.3 Computer Simulations	69
2.4 Raw Data Simulation	76
2.5 Perspective On Pairwise Strategies	78
2.6 Mathematical Models	80
2.7 Discussion	81
CHAPTER 3: VERTEBRATE TRYPSINOGEN EVOLUTION	86
3.1 Coincidental Evolution	87
3.2 Historical Perspective On Trypsinogen	89
3.3 The Compilation Of Trypsin And Trypsinogen Sequences	91
3.4 Cloning And Sequencing <i>Petromyzon marinus</i> Trypsinogen	94
3.5 Lamprey Trypsinogens	101
3.6 Cloning And Sequencing <i>Boltenia villosa</i> Trypsinogen	102
3.7 Tunicate Trypsinogen	103
3.8 Signal Sequences	104
3.9 Activation Peptides	107
3.10 Cystine Bridges	108
3.11 Insertions And Deletions	111
3.12 Intron/Exon Boundaries	112
3.13 Cationic And Anionic Trypsins	116
3.14 Amino-Acid Composition	118
3.15 Multiple Sequence Alignments	127

3.16 Sequence Distances	129
3.17 Multidimensional Scaling	132
3.18 Phylogenies	145
3.19 Modes Of Trypsinogen Evolution	167
3.20 Expression Of Trypsin	170
3.21 Function Of Trypsin	174
3.22 Genesis Of Novel Genes	176
BIBLIOGRAPHY	179
APPENDIX A: DOUBLE-BARREL SHOTGUN ALGEBRA	205
A.1 One Clone	206
A.2 Two Clones	206
A.3 Three Clones	208
A.4 More Than Three Clones	217
APPENDIX B: ALTERNATIVE SEQUENCE DISTANCE METRICS	221
B.1 Sequence Distances	221
B.2 The Jukes-Cantor Model	224
B.3 A Peptide View of The Jukes-Cantor Model	225
APPENDIX C: PCR EVALUATION OF TRYPSINOGEN EXPRESSION	236
C.1 Methods and Reults	236
C.2 Discussion	237

LIST OF FIGURES

<i>Figure</i>	<i>Page</i>
Figure 1.1. Schematic cartoon of a random subcloning project.	17
Figure 1.2. Schematic of a mathematical formulation for random subcloning.	22
Figure 1.3. Expected target coverage with respect to redundancy.	25
Figure 1.4. Relative error of the Lander-Waterman and “beta” models.	42
Figure 1.5. Computer simulations.	52
Figure 1.6. Expected closure costs.	56
Figure 1.7. Expected cost of reaching gapped project states.	59
Figure 1.8. Probability of completion with respect to redundancy.	61
Figure 1.9. Incremental cost of closing one gap.	63
Figure 2.1. Model “double-barrel shotgun” assembly.	67
Figure 2.2. Parameters from a 35 kb pairwise project with respect to redundancy.	72
Figure 2.3. Parameters from a 200 kb pairwise project with respect to redundancy.	73
Figure 2.4. Simulation of pairwise strategies employing a mix of insert sizes.	74
Figure 2.5. Simulation of a hybrid pairwise strategy.	75
Figure 3.1. Cartoons of genomic trypsinogen organization.	92
Figure 3.2. Stereoscopic view of the trypsin backbone.	108
Figure 3.3. Multiple alignment of chordate trypsin protein sequences.	113
Figure 3.4. Principal-component analysis of composition.	124
Figure 3.5. Multidimensionally scaled vertebrate trypsin sequence distances.	134
Figure 3.6. Stress vs. Dimensions.	136
Figure 3.7. Group I vs. Group II.	137
Figure 3.8. Allopatric division of the trypsinogen multigene family.	138
Figure 3.9. Group I vs. Group II (3-D).	139
Figure 3.10. Group I vs. Group II (alternative 3-D).	140
Figure 3.11. 5' vs. 3'.	141
Figure 3.12. Anionic vs. Cationic.	142
Figure 3.13. Multidimensionally scaled projection of the rodent group I trypsins.	144
Figure 3.14. Hypothetical phylogeny of the vertebrate trypsins.	146
Figure 3.15. Fitch-Margoliash phylogeny of forty-two vertebrate trypsins.	147
Figure 3.16. Fitch-Margoliash phylogeny of thirty-two vertebrate trypsins.	148
Figure 3.17. Weighted covariance statistics of thirty sequences.	150
Figure 3.18. Unweighted covariance statistics of thirty sequences.	152
Figure 3.19. Pseudogenes added to the vertebrate trypsin phylogeny.	153
Figure 3.20. Skewed Fitch-Margoliash phylogeny of the vertebrate trypsins.	155
Figure 3.21. A phylogeny of the rodent group I trypsins.	156
Figure 3.22. Alignment of newly sequenced trypsins.	158
Figure 3.23. Phylogeny of chordate trypsins.	161
Figure 3.24. Phylogeny of serine proteases.	165
Figure A.1. Topologies for one- and two-clone double-barrel configurations.	206
Figure A.2. Topologies for three-clone double-barrel configurations.	207
Figure A.3. Probability of one scaffold with respect to insert:fragment ratio.	218
Figure B.1. Trypsin site variability.	230
Figure B.2. Trypsin sequence logo.	231
Figure B.3. An alternative Fitch-Margoliash trypsin phylogeny.	233

LIST OF TABLES

<i>Table</i>	<i>Page</i>
Table 2.1. Results from a raw data simulation of a pairwise strategy.	78
Table 3.1. Literature references for the chordate trypsinogens.	95
Table 3.2. Classification comments for the chordate trypsinogens.	98
Table 3.3. PCR Primers used in the analysis of the chordate trypsinogens.	100
Table 3.4. Signal peptides and activation peptides of the chordate trypsinogens.	105
Table 3.5. Cystine bridges of the chordate trypsinogens.	110
Table 3.6. Predicted isoelectric points and charges of the chordate trypsins.	117
Table 3.7. Amino-acid compositions (sequenced).	120
Table 3.8. Amino-acid compositions (biochemical).	121
Table 3.9. Literature references for trypsin amino-acid compositions.	123
Table 3.10. Psychrophilic trypsin residue changes.	163
Table 3.11. Genbank identification numbers for the human trypsinogen ESTs.	173
Table 3.12. Genbank accession numbers for the mouse trypsinogen ESTs.	175
Table C.1. Number of cDNAs sequenced by PCR from several human tissues.	237
Table C.2. Number of cDNAs sequenced by PCR from several mouse tissues.	238

GLOSSARY

Coincidental Evolution. A tendency of genes present in the same genome to evolve in a non-independent manner. The presence of coincidental evolution makes homologous genes within a genome more similar to each other than to homologous genes from other genomes. Many authors use the term “concerted evolution” as a synonym for coincidental evolution, which was originally defined by Hood et al. (1975).

Contig. An island consisting of at least two fragments.

Fitness. An organism’s ability to propagate.

Functional Genomics. The development and application of experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics (Heiter and Boguski, 1997; McKusick, 1997).

Gap. A region of a target that is not represented in an island. Gaps are sometimes referred to as “oceans.”

Gene Product. The product of a gene. Most genes code for proteins, but some code for structural RNAs, and some affect the structure and/or regulation of the genome without being transcribed into a downstream product.

Gene. Information encoded in a segment of genomic DNA that affects the fitness of an organism. Semantically, it is useful to consider some genes as consisting of multiple gene segments, such as the TCR gene.

Genome. The information encoded in the DNA of a cell. Every individual, with few exceptions, has a distinct genome. The genome can vary slightly from cell to cell within an organism. The term was coined in 1920 by Winkler as an elision of the words “gene” and “chromosome” (McKusick, 1997).

Genomicist. One who studies genomes.

Genomics. The study of genomes. By contrast, the term “genetics” refers to the study of inheritance. The term “genomics” was introduced by Roderick in 1986 (McKusick, 1997).

Homologous Genes. Genes that share a common ancestral gene. Orthology and paralogy are subcategories of homology. Genes may be homologous without necessarily being either paralogous or orthologous (see, for example, Tatusov et al., 1997).

Indel. An insertion or a deletion.

Island. A maximal set of fragments each of which is connected to all other island members by at least one path of overlapping fragments.

Isozymes. Enzymes that have identical (or nearly identical) biochemical properties.

Orphon. A term applied to gene segments separated from a complete functional gene locus. A typical example is a TCR V segment on chromosome 9, unable to recombine with D and C segments to form a functional gene.

Orthologous Genes. Genes in different species that share a common function and evolved from a common ancestor. By definition, the split between such genes was caused by a speciation event. The result of speciation (Fitch, 1970).

Paralogous Genes. Multiple genes resulting from duplication within a particular genome. The result of gene duplication (Fitch, 1970).

Parameterization. A particular choice of parameters for a model (or a project). Parameters are variables that one can control, such as the choice of how many clones to analyze, or what length clones to choose.

Pseudogene. A gene that is no longer functional. The ancestral sequence was a functional gene that acquired one or more mutations that destroyed functionality. Typically, this would entail the acquisition of stop codons in an open reading frame. Point mutations are typically responsible for the creation of pseudogenes.

Relic. A fragment of a gene that was once functional. The semantic boundary between a pseudogene and a relic is fuzzy. Generally in order to be classified a relic, one or more recombinations or major deletions must have operated on the ancestral gene.

Repeat. A region of a genome that is nearly identical to another region of the same genome. It should be emphasized that in genomics terminology the word “repeat” does not imply 100% identity. The percent identity used to define a repeat is somewhat subjective.

Scaffold. An ordered and oriented list of islands. Also referred to as a “gapped island” (Port et al., 1995) or a “supercontig” (Lawrence et al., 1994).

Structural Genomics. The construction of genetic, physical, and transcript maps of genomes (Heiter and Boguski, 1997). Genomic sequence is considered to be the ultimate high resolution physical map of a genome. The definition can be rephrased as, “mapping and sequencing genomes” (McKusick, 1997). This original definition has not been replaced in common usage. The term “structural genomics” now usually refers to the elucidation of all protein structures coded by a genome. This new meaning would be better served by the term “structural proteomics”, but this phrase is unfortunately not in usage.

Subclone. A clone of a fragment of a larger piece of DNA. The fragment has been genetically engineered into a vector that facilitates laboratory manipulation of the fragment.

Target. A genome or a subset of a genome that will be analyzed during the course of a project.

LIST OF ABBREVIATIONS

BAC. Bacterial artificial chromosome.

bp. Base pair.

cDNA. Complementary deoxyribonucleic acid.

DNA. Deoxyribonucleic acid.

ds. Double stranded.

EST. Expressed sequence tag.

HMM. Hidden Markov model.

kb. Kilobase pair.

mRNA. Messenger RNA.

OSS. Ordered shotgun sequencing.

PCR. Polymerase chain reaction.

RNA. Ribonucleic acid.

SMG. Sequence-mapped gap.

ss. Single stranded.

STS. Sequence-tagged connector.

STS. Sequence-tagged site.

TCR. T-cell receptor.

YAC. Yeast artificial chromosome.

LIST OF VARIABLES

C . An arbitrary length (a partial target goal less than G).

D_k . The length of the k^{th} spacing from the left end of the target.

f . The fraction of the target that is covered.

f_g . The effective fractional coverage of the target provided by one fragment.

G . The length of a target (in bases). This abbreviation is most appropriate when the target is a genome, but has come to be used even when the target is something else, such as a subclone. I do not employ the convention of setting G equal to unity.

G_e . The effective length of the target.

g_m . The number of short spacings in the m^{th} island from the left end of the target.

I . The length of an insert.

L . The length of a clone (in bases). I do not employ the convention of setting L equal to unity.

l_m . The length of the m^{th} island from the left end of the target.

N . The number of permitted residues at a sequence site.

n . The total number of clones analyzed in a project, either by mapping or sequencing.

N_{gaps} . The number of gaps in a project.

N_{islands} . The number of islands in a project.

N_{long} . The number of long spacings in a project.

$N_{\text{singletons}}$. The number of single fragment islands in a project.

μ . The probability of a mutation per unit evolutionary time.

p_{gap} . The probability of a gap following a spacing.

R . Redundancy.

R_e . The effective redundancy.

S_k . The k^{th} spacing from the left end of the target.

. Evolutionary time.

T. The number of base pairs necessary to determine overlap during the assembly phase of a random subcloning project. In many practical cases $T \ll L$, and can be approximated as zero.

V. The length of the vector sequence.

z_m. The number of fragments in the m^{th} island from the left end of the target.

PREFACE

The doctoral thesis has a rich history. The tradition of the dissertation binds modern students back through the ages to the history and culture of the Renaissance Universities (Haskins, 1923). The richness of values governing the content of theses dictates certain compromises with respect to the choice of included material and its style of presentation.

My desire in this thesis is to provide knowledge in a understandable fashion to the widest possible audience: biologists, mathematicians, computer scientists, engineers, and those in allied fields. I hope to have made at least a portion of the text accessible to students at all levels. However, for much of this dissertation, I will assume at least a basic knowledge of molecular biology, such as that obtainable from *The Cartoon Guide to Genetics* (Gonick and Wheelis, 1991) or another introductory molecular biology textbook. The reader is encouraged to ignore or merely skim material that is either too basic or advanced. A more compact presentation of much of the material in this dissertation can be found in my original journal articles which are reproduced in the appendices.

My doctoral research has led me down many disparate paths, of which three have been significant enough to include in this dissertation, each as a separate chapter. Two of these chapters group naturally in the area of strategic genomics, while the third strikes out tangentially into the field of molecular evolution. Rather than attempt an integrated and detailed introduction to these three subjects simultaneously, I have kept the overall introduction, which follows, short and generally oriented. I include more detailed topical introductions at the beginning of each chapter. Also, in each chapter I employ several specialized terms appropriate to the subject matter; I encourage the reader to exploit the glossary when in doubt of the meaning of a particular term. Readers may also find the list of abbreviations and the list of variables useful as occasional references.

This thesis is a living document. The most recent edition of this thesis can be found on the web. My current website is faculty.washington.edu/roach. Errata notifications and suggestions for additions or improvements are appreciated.

ACKNOWLEDGEMENTS

Collaboration is the cornerstone of modern research. One can lift a rock; many can move a mountain. The many include: Lee Hood, Roger Perlmutter, Chris Wilson, Dave Lewis, Ken Walsh, Andy Siegel, Steve Henikoff, Kai Wang, and innumerable technicians, graduate students, post-docs, professors, friends, and family, all who have contributed variously their labor, ideas, and encouragement. I am grateful for a grant from the Life & Health Insurance Medical Research Fund.