Computational Biology

Notes by: Anthony Colombo
Harvard University

Statistics - Spring 2016
Professor: Dr. Shirley Liu

Updated May 1, 2016

Contents

Definitions

Theorems, Lemmas, and Corollaries

Introduction

This set of notes was taken by hand by Anthony Colombo in his *Computational Biology* course at Harvard. They were later typed and prepared for typesetting using LaTeX.

Organization is as follows:

- The notes were gained from Dr. Liu's lectures and lab-work.
- Definitions are numbered according to the section in which they are defined. These numbers do not necessarily correspond to the text.
- When there would otherwise be ambiguity, page numbers/references to the text will be added.

1 Notation

Definition 1.1 (Notation).

- *1: we will always denote $\mathbb{R}$ as a commutative ring with 1, this identifies the identity element as 1 in the non-empty set R that satisfies all the field operations.*
- *2: We will use $\mathbb{F}$ to denote a Field.*
- *3: We will use R [x]: $\{a_0 + a_1x + ... + a_kx^k | k \in \mathbb{N} \forall a_j \in R\}$ will be a commutative ring with 1.*
- *4: For any prime p and $m \in \mathbb{N}$ $\exists$ a field $\mathbb{F}_{|}$ that has p-1 elements denoted $\{1, 2, ..., p-1\}$.*
- *5: the notation for matrices is as follows : for any R (a commutative ring with 1) and $\forall m, n \in \mathbb{N}$ R^{nXm} is the set of all n X m matrices over R. so $A \in R^{nXm} \iff A =$*

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{nm} \end{pmatrix}$$

- *6: for R^{nXn} it is denoted as $M_n(R)$, ie. all the n X n matrices that operate under the field operations and are commutative rings with 1.*
- *7: The n X n identity matrix is denoted as I_n =Diag(1,...,1) and the zero matrix of an n X m matrix that has all entries as 0 is denoted as 0_{nXm}.*
- *8: the unit matrix is defined as with zeroes in all the entries except the position i,j denoted as e_{ij}.*

2 Linear Equations

Definition 2.1 (Field Notation).

- *1: denote $\mathbb{F}$ as either the set of real numbers or the set of complex numbers, and $\mathbb{F}$ is also a field.*

Definition 2.2 (Field Properties). *A field must have the following properties*

- *1: a field is an abelian group under* ($\mathbb{F}$, +) *such that under addition* $\forall$ *x,f* $\in \mathbb{F}$ *addition is commutative, ie x+y = y+x*
- *2: a field is associative, x+(y+z) = (x+y)+z* $\forall$ *x,y,z* $\in \mathbb{F}$.
- *3:* $\exists$ *an identity 0 under addition such that x+0 = x*
- *4:* $\exists$ *an inverse under addition such that* $\forall$ $x \in \mathbb{F}$ x^{-1} *=-x, where x+(-x) = 0 (the identity)*
- *5: A field is an abelian group under multiplication such that* $\forall$ *x,y* $\in \mathbb{F}$ *xy =yx*
- *6: Under multiplication, the operation is associative ie.* $\forall$ *x,y,z* $\in \mathbb{F}$ *x(yz)=(xy)z.*
- *7: There exists an identity 1 such that x1=1x=x.*
- *8: There exists an inverse such that* xx^{-1}=1 *where* $x^{-1}=\frac{1}{x}$
- *9: Multiplication operations distributes over addition, ie. x(y+z)=xy+xz*

Definition 2.3 (Characteristic). *the smallest n where n* $\in \mathbb{F}$ *such that the sum of n (could be 1) is equal to 0 is called the characteristic. ie.* $\sum$ *1 = 0*

Definition 2.4 (Characteristic Zero). *the characteristic Zero is where there is no smallest element n, such that the sum of n's is equal to zero.*

2.1 Linear Equations

Definition 2.5 (System of Equations). *Suppose $\mathbb{F}$ is a field. we consider the problem of finding n scalars $(x_1, x_2, ..., x_n)$ which satisfy the conditions: Consider this equation (1-1)*
$A_{11}x_1 + ... + A_{1n}x_n = y_1$
$A_{21}x_1 + ... + A_{2n}x_n = y_2$
. . . .
. . . .
$A_{m1}x_1 + ... + A_{mn}x_n = y_m$ *where the n-tuple $(x_1, x_2, ..., x_n)$ of $\mathbb{F}$ satisfy the system of equations with n unknowns; this n-tuple is called the solution. if each $\forall$ i y_i =0 for each of the n equations the system of equations is called homogenous.*

Definition 2.6 (Linear Combination). *now suppose we have m scalars $(c_1, c_2, ..., c_m)$ and we multiply the j^{th} system in (1-1) by the scalar c_j, and then add the results column wise, combining each of the coefficients under the variable x_j. This will result in an equation:*
$(c_1A_{11} + ... + c_mA_{m1})x_1$+...+$(c_1A_{1n} + ... + c_mA_{mn})x_n$=$c_1y_1$ + ...c_my_m
important: *any solution to the system of the linear combinations is a solution to the original system (1-1), these systems are equivalent*

Definition 2.7 (Equivalent, Row equivalent). *Suppose we have a new system of equations defined as : $B_{11}x_1 + ... + B_{1n}x_n = z_1$*
$B_{21}x_1 + ... + B_{2n}x_n = z_2$
. . . .
. . . .
$B_{k1}x_1 + ... + B_{kn}x_n = z_k$ *If this system was derived from the linear combinations of (1-1), then every solution of (1-1) is a solution to this system. Thus two systems are defined at* ***equivalent*** *if each system is a linear combination of the other system.*

Theorem 2.1 (Equivalent Systems of Equations). *Equivalent systems of linear equations have exactly the same solutions*

2.2 Matrices and Elementary Row Operations

Definition 2.8 (Notation on Matrices). *the system of equations in (1-1) can be written in matrix form such that AX=Y is defined as A, the matrix of coefficients is :*

$$AX = Y \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ ...x_k \\ ...x_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_k \\ ...Y_n \end{pmatrix}$$

Here A is the matrix of coefficients , X is a vector of variables and Y is a vector of solutions.

Definition 2.9 (Elementary Row Operations). *Now considering A in the equations AX=Y, we are considering the rows of A which can be used to form linear combinations using elemenatary row operations on an m X n matrix A over a field F using:*

- *1: multiplication of one row of A by a non-zero scalar c.* $e(A)_{ij} = A_{ij}$, *if* $i \neq r$, $e(A_{rj}) = cA_{rj}$.
- *2: replacement of the* r^{th} *row in A by row r plus c times row s, c is any scalar and* $r \neq s$. $e(A_{ij}) = A_{ij}$, *if if* $i \neq r$ *then,* $e(A_{rj}) = A_{rj} + cA_{sj}$.
- *3: interchanging any two different rows in A.* $e(A_{ij}) = A_{ij}$, *if* $i \neq r \neq s$, *then* $e(A_{rj}) = A_{sj}$, *and* $e(A_{sj}) = A_{rj}$

Theorem 2.2 (Inverse Operations of Elementary Row Operations). *Elementary operations* $e_{1,2,3}$ *are considered to be functions, along with the matrix operations such as transpose. Thus the elementary row operations have inverse functions and are well defined. To each elementary row operation* e_i *there corresponds an elementary row operation* e_1 *such that* e_1 *is the same type as e and* $e_1(e(A)) = e(e_1(A))$ $= A$ *for each A. In other words, there exists inverse row operation (or function) and is the same type as the elementary row operation.*

Definition 2.10 (Row-Equivalent). *if A and B are m X n matrices over the field F, we say that A is row equivalent to B if A can be formed from row operations of B.*

Theorem 2.3 (Solutions to Row Equivalent Matrices of Homogenous Equations). *If A and B are m X n matrices that are row equivalent and AX=0 and BX=0 then they have the same exact solutions.*

Definition 2.11 (Row Reduced).

- *1: each row has a leading one entry, where leading one, is the first non-zero entry*
- *2: the columns of an m X n matrix which contain the leading one have all other entries as 0 for the other rows.*

Theorem 2.4 (Matrices in a Field and Row Reduced Matrices). *Every matrix in a field , F, has a row reduced matrix. In other words, there are no matrices that can not be row reduced.*

Definition 2.12 (Reduced Row Echelon Form). *let R be an m X n matrix over a field F.*

- *1: R must be row reduced*
- *2: all the rows in which have only zero entries occur below rows that have non-zero entries*
- *3: if the rows 1,...r, are the non-zero entries of R, then for k_i as the column that has the leading one, k_1¡ ... ¡k_r. (ie. order of appearance is from left to right)*

Theorem 2.5 (Row Equivalence and Reduce Row Echelon Matrices). *Every m X n matrix in a field F is row equivalent to a reduced row echelon matrix.*

Theorem 2.6 (Non- Trivial Solutions and Homogenous Equations). *If A is an m X n matrix and m ¡ n then the homogenous solution contains a non-trivial solution to the equation AX=0.*
This can be understood as after row reducing a matrix, let there be r non-zero leading one rows and (n-r) uknowns. then we can assign any value to the columns that do not have a leading one. so if we have r non-zero rows, then we will have r non trivial equations, with r unknowns and (n-r) unknowns from the other columns. we can assign any value to the (n-r) unknowns.

Definition 2.13 (Augmented Matrix). *An augmented matrix is formed from the equation AX=Y and the column vector Y is augmented onto the coefficient matrix A to form a system of homogenous solutions. thus A is an m X n matrix, transforms into an m X (n+1) matrix. where*
$A'_{ij} = A_{ij}$
$A'_{i(n+1)} = y_i$

2.3 Multiplication of Matrices

Definition 2.14 (Matrix Multiplication). *Let A be an m X n matrix over the field F and let B be an n X p matrix over F, then the product AB is an m X p matrix over F, C, which is defined by entries i, j by* $C_{ij} = \sum_{r=1}^{n} A_{ir}B_{rj}$

2.4 Invertible Matrices

Definition 2.15 (Elementary Matrix). *an elementary matrix is an m X n matrix which can be obtained from an m X m identity matrix from a single row operation. Thus E= e(I)*

Theorem 2.7 (Elementary Matrices and Non-Square Matrices). *Let e be an elementary row operation and let E be the square m X m elementary matrix E = e(I), then for ever m X n matrix A e(A)=EA.*

Definition 2.16 (Left and Right Inverses). *let A be an n X n square matrix and let B be an n X n square matrix, then the left inverse is defind as BA = I, and the right inverse is defined as AB = I. If AB=BA=I then B is called the two sided inverse, and A is called an invertible matrix*

Theorem 2.8 (Square Matrices and Properties of Inverses). *Let A and B be n X n matrices*

- *1: If A is invertible, then so is* A^{-1}, *where* $(A^{-1})^{-1} = A$
- *2: If A and B are both invertible (ie. have two sided inverses) then* $(AB)^{-1} = B^{-1}A^{-1}$

Theorem 2.9 (Products of Invertible Matrices). *Products of Invertible matrices are invertible*

Theorem 2.10 (Elementary Matrices and Inverses). *Elementary matrices are invertible*

Theorem 2.11 (Square Matrices and Invertible Properties). *Let A be an n X n matrix then :*

- *1: A is invertible*
- *2: A is row - equivalent to the n X n identity matrix*
- *3: A is a product of elementary row operations*

Theorem 2.12 (Square Matrix With Left or Right Inverse). *A square matrix with either a left or right inverse is invertible.*

3 Lecture Notes from Dr. Liu

The introduction to Sequencing studies in biology dates back to the study of protein blocks and predictions studies from sequencing amino acids.

- 1955: Sanger sequenced the first protein bovine insulin
- 1970: Smith-Waterman algorithm developed, alignment to identify shared commonalities of amino acid sequences to determine function predictions amidst groups
- 1973: protein databases created Protein Data Bank.
- 1990: BLAST tool created by Broad (?) to determine novel transcripts
- 1994 : BLOCKS protein shape derived from shared amino acids sequence commonalities, shape predictions, function predictions.
- 1994: CASP crystals are difficult to make, where de-novo validation is done through crystallography. CASP identifies tertiary structure, and/or homology from sequences known as templates. A template is defined as a protein sequence in terms of amino acids of known functionality.
- 1997: Proteomics
- 1995: MicroArray. Gene chips formed in the Bay area arise, cost roughly $ 1300 per chip; the need arose from limitations of western-blots and qt-PCR which quantify one gene at a time. MicroArray has multiple probes where RNA will hybridize to a set of probes. Each probe can measure different features

3.1 DNA Sequencing

- 1953: DNA Seq structure found by Watson and Crick, who used supportive data from Rosalind Franklin as a proof of concept. Rosalind Franklin specialized in X-Ray crystallography who died of abdomen cancer likely to have arose from playing with X-Rays. She photographed the DNA double helix, and her boss Walkins showed the photo to Crick who immediately understood the meaning. However Franklin did not mathematically derive the angles between the helix, and did not prove the chemistry behind this image; her data was used but also uncited giving rise to controversy of women in academia (sexism).
- 1972: recombinant DNA and ligase techniques
- 1977: Sanger DNA sequencing, this works for large volumes
- 1985: qt PCR, library amplification
- 1988: NCBI records any and all data, BLAST for novel transcripts
- 1990: BLAST
- 1990-2003: Human Genome Project the complete assembly of 23 pairs of chromosomes. where each chromosome was assembled across different institutions and private businesses. The sequencing limits existed for at most 1000 bp, so they fragmented each chromosome into sub-encyclopedias, or sub-libraries, and sequenced them. they used

a computer algorithm from a PhD Computer Scientist at UC Davis to assemble the fragmented sequences.

The real essence of this field begins with computer science, but ends with Biology. Biology allows for the proper question in a field with infintely abundant data.

3.2 Micro Array

MicroArray (MA) is a gene chip with many different cells