Advanced Algorithms and Models for Computational Biology

-- a machine learning approach

Population Genetics:

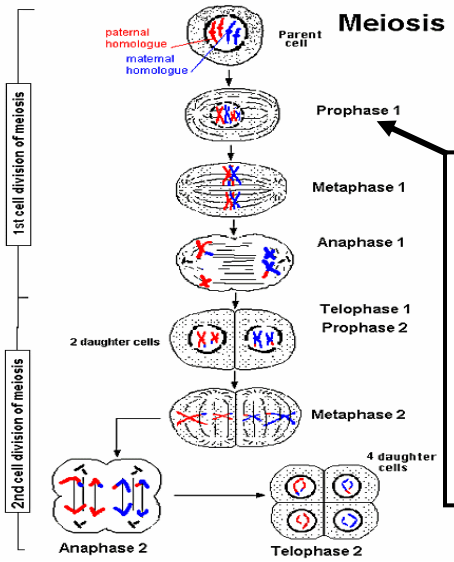
meiosis and recombination

Eric Xing

Lecture 15, March 8, 2006

Reading: DTW book,

Meiosis



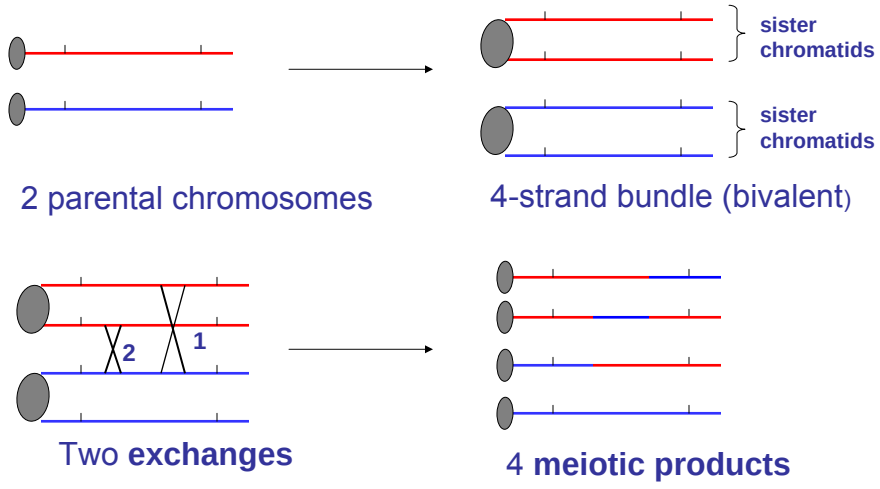
Meiosis is a process which starts with

The action of interest to us happens around here :

- Chromosomes **replicate**, but stay joined at their **centromeres**
- **Bivalents** form
- **Chiasmata** appear
- Bivalents separate by attachment of centromeres to **spindles**.

Source: http://www.accessexcellence.org

Four-strand bundle and exchanges



Chance aspects of meiosis

- Number of exchanges along the 4-strand bundle
- Positions of the exchanges
- Strands involved in the exchanges
- Spindle-centromere attachment at the 1st meiotic division
- Spindle-centromere attachment at the 2nd meiotic division
- Sampling of meiotic products

Deviations from randomness are called ***interference.***

A stochastic model for meiosis

- A point process X for exchanges along the 4-strand bundle
- A model for determining strand involvement in exchanges
- A model for determining the outcomes of spindle-centromere attachments at both meiotic divisions
- A sampling model for meiotic products

Random at all stages defines the ***no-interference*** or ***Poisson*** model.

A model for strand involvement

- The standard assumption here is

No Chromatid Interference (NCI):

each **non-sister** pair of chromatids is equally likely to be involved in each exchange, independently of the strands involved in other exchanges.

NCI fits pretty well, but there are broader models.

Changes of parental origin along *meiotic products* are called ***crossovers***. They form the crossover point process *C* along the single chromosomes.

Under NCI, *C* is a Bernoulli *thinning* of *X* with *p=0.5*.

From exchanges to crossovers

- Usually we can't observe exchanges, but on suitably marked chromosomes we can track crossovers.

  Call a meiotic product ***recombinant*** across an interval *J*, and write *R(J)*, if the parental origins of its endpoints differ, i.e. if an ***odd number of crossovers*** have occurred along *J*. Assays exist for determining whether this is so.

- **Mather's formular:**
  Under NCI we find that if $n>0$, $pr(R(J) \mid X(J) = n) = 1/2$, so

  $$pr(R(J)) = 1/2 \times pr(X(J) > 0) \ldots\ldots\ldots (*)$$ (Proof?)

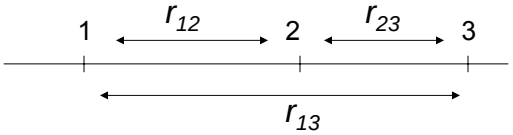
Recombination and mapping

The ***recombination fraction*** *pr(R(J))* gives an indication of the chromosomal length of the interval *J*: under NCI, it is monotone in |J|.

Sturtevant (1913) first used recombination fractions to order (i.e. **map**) genes. (How?)

Problem: the recombination fraction does not define a metric.

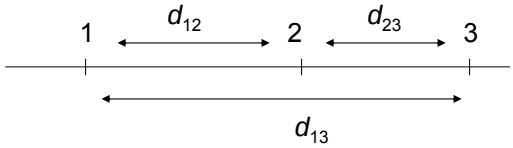
Put $r_{ij} = pr(R(i--j))$.



$$r_{13} \neq r_{12} + r_{23}$$

Map distance and mapping

- **Map distance**: d_{12} = E{C(1--2)} = av # COs in 1--2
  - ***Unit***: Morgan, or centiMorgan.



$d_{13} = d_{12} + d_{23}$ (expectations are additive!)

  - The expectation says nothing definitive about the relationship **physical** distance and **genetic** distance
- **Genetic mapping** or applied meiosis: a **BIG** business
  - Placing genes and other markers along chromosomes;
  - Ordering them in relation to one another;
  - Assigning map distances to pairs, and then globally.

Genetic linkage

Haldane's model:

These crossovers occur as a Poisson process of rate d (per Morgan).

$$\rho(d) = \sum_{k \text{ odd}} e^{-d} \frac{d^k}{k!} = \frac{1}{2} e^{-d} \sum_{k=0}^{\infty} \left(\frac{d^k}{k!} - \frac{(-d)^k}{k!} \right)$$

$$= \frac{1}{2}(1 - \exp(-2d)).$$

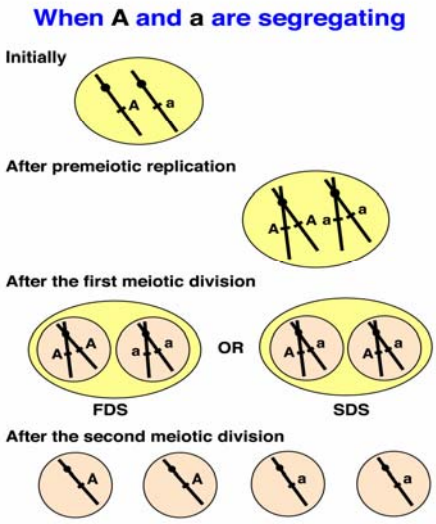
ρ(d) is an increasing function of d, ρ(d) → 1/2 as d → ∞, and ρ(d) ≈ d as d → 0.

The program from now on

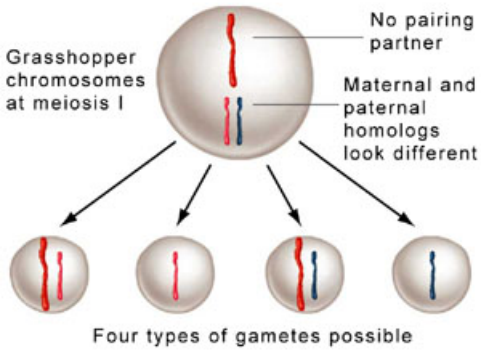
- With these preliminaries, we turn now to the data and models in the literature which throw light on the chance aspects of meiosis.
- **Mendel's law of segregation**: a result of random sampling of meiotic products, with allele (variant) pairs generally segregating in precisely equal numbers.

As usual in biology, there are exceptions.

Segregation



Random spindle-centromere attachment at 1st meiotic division



In 300 meioses in an grasshopper heterozygous for an inequality in the size of one of its chromosomes, the smaller of the two chromosomes moved with the single X 146 times, while the larger did so 154 times.

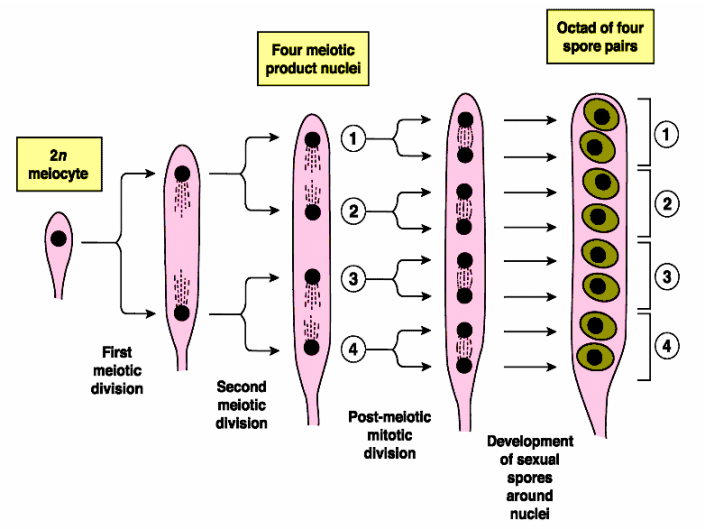
Carothers, 1913.

Tetrads

- In some organisms - fungi, molds, yeasts - all four products of an individual meiosis can be recovered together in what is known as an *ascus.* These are called ***tetrads***. The four *ascospores* can be typed individually.

- In some cases - e.g. *N. crassa*, the red bread mold - there has been one further mitotic division, but the resulting ***octads*** are *ordered.*

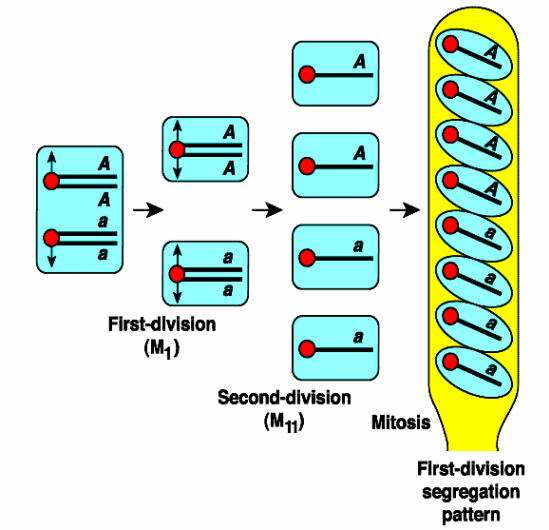
Meiosis in N.crassa



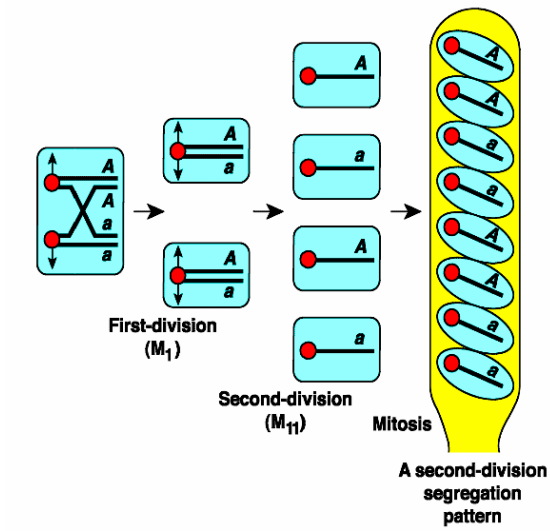
Using ordered tetrads to study meiosis

- Data from ordered tetrads tell us a lot about meiosis. For example, we can see clear evidence of 1st and 2nd division segregation.
- We first learned definitively that normal exchanges occur at the 4-stand stage using data from *N. crassa*, and we can also see that random spindle-centromere attachment is the case for this organism.
- Finally, aberrant segregations can occasionally be observed in octads.

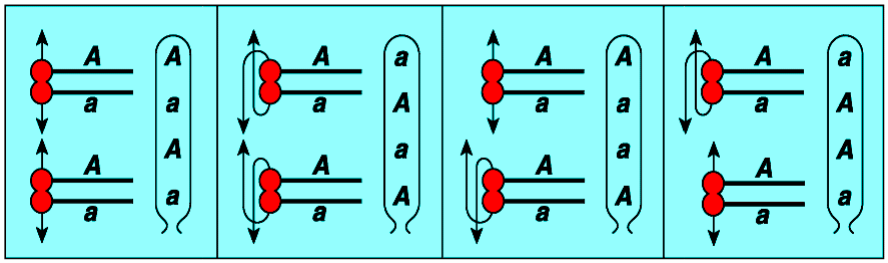
First-division segregation patterns



Second-division segregation patterns



Different 2nd division segregation patterns



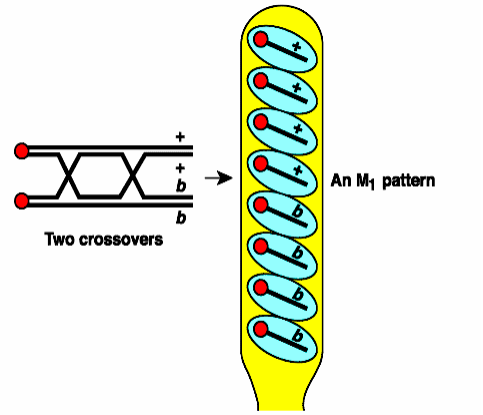
Under random spindle-centromere attachment, all four patterns should be equally frequent.

Lindegren's 1932 N. crassa data

Class	Position of spore in ascus								Number
	1	2	3	4	5	6	7	8	
I	A A	A A	A A	A a	a A	a a	a a	a a	102 3 } 105
II	a a	a a	a a	a A	A a	A A	A A	A A	123 6 } 129
III	A A	A a	a A	a a	A A	A A	a a	a a	8 1 } 9
IV	a	a	A	A	a	a	A	A	5
V	A A	A a	a A	a a	a a	a a	A A	A A	10 1 } 11
VI	a	a	A	A	A	A	a	a	14
								Total	273

2-strand double exchanges lead to FDS

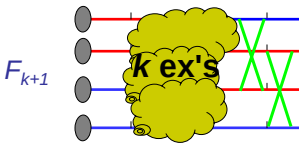
There is a nice connection between the frequencies of multiple exchanges between a locus and its centromere and the frequency of 2nd division segregations at that locus.



A simple calculation and result

- Let F_k (resp. S_k) denote the number of strand-choice configurations for k exchanges leading to *first* (resp. *second*) division segregation at a segregating locus. By simple counting we find

$F_0 = 1$ and $S_o = 0$, while for $k>0$,
$F_{k+1} = 2S_k$, and $S_{k+1} = 4F_k + 2S_k$.



S_{k+1}: ? (homework)

Map function from SDS

- Assuming NCI, the proportion S_k of second-division segregants among meioses having k exchanges **between our locus and the centromere** is

$$s_k = \frac{2}{3}[1-(-\frac{1}{2})^k], \quad k > 0.$$

- If the distribution of the # of exchanges is (x_k), then the frequency of SDSs is

$$s = x_1 + \frac{1}{2}x_2 + \frac{3}{4}x_3 + ...$$

  - If the distribution is Poisson (2d) then we find

$$s = \frac{2}{3}(1-e^{-3d}).$$

  - This is a **map-function**: between the unobservable map distance *d* and the observable SDS frequency *s*.

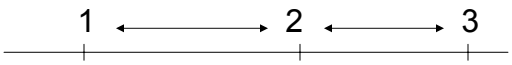
Interference: the state of play

- Total number of exchanges on an arm rarely Poisson
- Positions of exchanges rarely Poisson in map distance (i.e. crossover interference is the norm)
- Strand involvement generally random (i.e. chromatid interference is rare)
- Spindle-centromere attachment generally random (non-random attachments are quite rare)

- The biological basis for crossover interference is only slowly becoming revealed; stay tuned.

Crossover interference

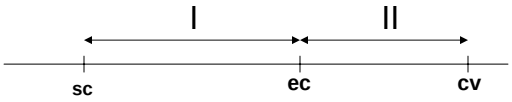
- The Poisson model implies independence of recombination across disjoint intervals



$$pr(R(1\text{--}2) \ \& \ R(2\text{--}3)) = pr(R(1\text{--}2)) \times pr(R(2\text{--}3))$$

Proof?

Morgan's D. melanogaster data (1935)



0: no recombination; 1: recombination

	0	1
0	13670	824
1	1636	6*

* the number of double recombinants that we would expect if recombination events across the two intervals were independent is 85

- Clearly there are many fewer double recombinants than the independence model would predict.
- This phenomenon is called **crossover interference**..

A measure of crossover interference

The **coincidence coefficient** S_4 for 1--2 & 3--4 is:

$$\frac{pr(R(1\text{--}2)\ \&\ R(3\text{--}4))}{pr(R(1\text{--}2)) \times pr(R(3\text{--}4))}$$

$$= \frac{pr(R(1\text{--}2) \mid R(3\text{--}4))}{pr(R(1\text{--}2))}$$

No crossover interference (for these intervals) if $S_4 = 1$
Positive interference (inhibition) if $S_4 < 1$.

Stochastic models for exchanges

- Count-location models
- Renewal process models
- Other special models, including a polymerization model

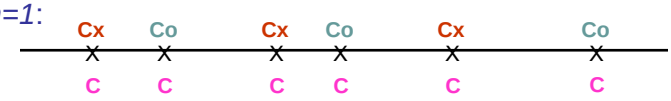
Count-Location Models

- These models recognize that interference influences distribution of the ***number of exchanges***, but fail to recognize that the ***distance between them*** is relevant to interference, which limits their usefulness.
- Let *N* = # *exchanges* along the bivalent.
  1. *Count distribution*: $q_n = P(N = n)$
  2. *Location distribution*: individual exchanges are located independently along the four-strand bundle according to some common distribution *F*.
- *Map distance* over *[a, b]* is $d = \lambda[F(b) - F(a)]/2$, where $\lambda = E(N)$.

Barrett *et al* (1954), Karlin & Liberman (1979) and Risch & Lange(1979)

The Chi-Square Model

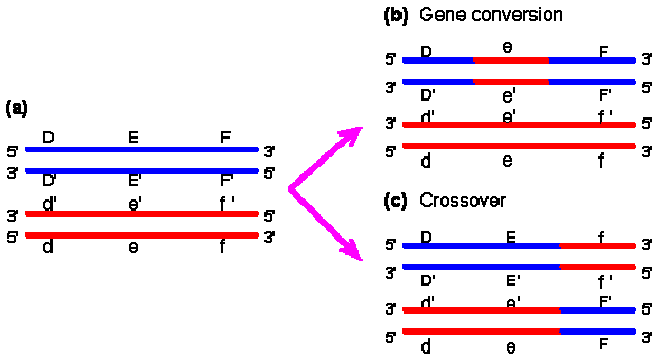
- Modeling exchanges along the 4-strand bundle as events from a ***stationary renewal process*** whose inter-event distribution is χ^2 with an even number of degrees of freedom. The *x* events are randomly distributed and every *(m+1)st* gives an exchange:
- *m=1*:



- The chi-square model is denoted by $Cx(Co)^m$.
  - *m* = 0 corresponds to the Poisson model.

Fisher *et al* (1947), Cobbs (1978), Stam (1979), Foss *et al* (1993), Zhao *et al* (1995)

Conversion vs. crossover



Biological interpretation of the chi-squared or *Cx*(*Co*)^m model

- The biological interpretation of the chi-squared model given in Foss, Lande, Stahl, and Steinberg 1993, is embodied in the notation *Cx*(*Co*)^m :

  The C events are crossover initiation events, and these resolve into either reciprocal exchange events Cx, or gene conversions Co, in a fairly regular way: crossovers are separated by an organism-specific number *m* of conversions.

Fitting the Chi-square Model to Various Organisms

Gamete data:

D. melanogaster:	$m = 4$
Mouse:	$m = 6$

Tetrad data:

N. crassa:	$m = 2$
S. cerevisiae:	$m = 0$ - 3 (mostly 1)
S. pombe:	$m = 0$

Pedigree data:

Human (CEPH):	$m = 4$

The chi-square model has been extremely successful in fitting data from a wide variety of organisms rather well.

Failure of the $Cx(Co)^m$ model with yeast

- The biological interpretation of the chi-squared model embodied in the notation $Cx(Co)^m$ is that crossovers are separated by an organism-specific number of potential conversion events without associated crossovers.
- *It predicts that close double crossovers should be enriched with conversion events that themselves are not associated with crossovers.*
- With yeast, this prediction can be tested with suitably marked chromosomes.

It was so tested in Foss and Stahl, 1995 and failed.

Challenges in the statistical study of meiosis

- Understanding the underlying biology
- Combinatorics: enumerating patterns
- Devising models for the observed phenomena
- Analysing single spore and tetrad data especially multilocus data
- Analysing crossover data

Acknowledgements

- Terry Speed, UC Berkeley
- Mary Sara McPeek, Chicago
- Hongyu Zhao, Yale
- Karl Broman, Johns Hopkins
- Franklin Stahl, Oregon

References

- **www.netspace.org/MendelWeb**
- HLK Whitehouse: **Towards an Understanding of the Mechanism of Heredity,** 3rd ed. 1973
- Kenneth Lange: **Mathematical and statistical methods for genetic analysis,** Springer 1997
- Elizabeth A Thompson **Statistical inference from genetic data on pedigrees,** CBMS, IMS, 2000**.**