Comparative Genomics
Homology/Paralogy/Orthology

Early Globin Gene

Gene duplication

α-globin

frog α  chick α  mouse α  frog β  chick β  mouse β

orthologs  paralogs

homologs
What is Genome Alignment

Find and align similar pieces of the genomes

Martin C. Frith
The human and chimp genomes are mostly very similar:

...actgctacgtacgtacggcatgctacgtagggcatgca-gctagcatgc...
   ||||||||x|||||||x||x||x||x||x||x||x||x
...actgcttcgtacgtacggcatgctacgtggggcatgcaggtagcatgc...

So why does this region have so many differences?

...cgtagctacgtacgtacgtagcattatcgtcgtctcgctatcgtgatcgt...
   ||||x||x||x||x||x||x||x||x||x||x||x
...cgtagccacgtaagttgcattatgtgtcgt----tgctatcgtgatcgt...

It may have a function that “makes us human”
Applications

The human and mouse genomes are mostly quite different:

```
...actgctacgtacgtacggcatgcta-----ggcatgcatgctagcatg...
```

```
||xx||x||xx|||xx|x||xxxxx|||x||x|x||xxx|xx||
```

```
...acaacttcgta--tacggggttctacgtggggcttgaagc---cgggc...
```

So why is this region so similar?

```
...cgtagctacgtacgtacgtagcattatcgtcgtatcgtgctatagtgatcgta...
```

```
|||xxxx|||xxx|||xxxx|xxxxx||x||xxx||xxxxx|||xxx||xxxxx|||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx||xxx|...
```

```
...cgtagctacgtacgtacgtagcattatcgtcgtatcgtgctatagtgatcgta---ctatagtgatcgta...
```

It may have a function essential for life in mammals
Applications

Human  ...actgctacgtacgtacggcatgctacgtaggg...
    |   |   |   |   |   |   |   |
Chicken ...accgcaacatacgtgcgccacgcgacttatgg...

Are these differences random, or is there a pattern?
Figure 1.4. Co-occurrence of genes or landmark sequences within single chromosomes or chromosome regions when chromosomes from each of two different organisms are compared. Panel A: Conserved synteny. In this case, \( g_{B1}, \ldots, g_{B3} \) represent genes in species B that have homologs \( g_{C1}, \ldots, g_{C3} \) in species C. Panel B: Syntenic segments and syntenic blocks. In this case, \( g_{B1}, \ldots, g_{B5} \) and the similar sequences in species C refer to landmark sequences on the genome, which can be more numerous than genes to produce a higher marker density. Syntenic segments are conceptually similar to conserved segments, except that there may be micro or rearrangements undetected because of the low marker density.

1.3.3 Biological String Manipulation

As indicated above, DNA is not immutable. During the copying or replication process, errors can occur (hopefully at low frequency, but at significantly high frequency in the case of reverse transcription of the HIV genome, for example).
Statistical Criteria

A. Genome Y
1 2 3 4 6 7
Genome X
1 2 3 4 5 6 7 8 9 10
collinear gene cluster

B. Genome Y
Genome X

C. Genome Y
Genome X

Mural et al., 2002
Genome Evolution
Genetic/Physical Map

List of genes/physical markers along a genome together with the distances between these genes/markers

- Mouse and human have almost the same set of genes but have completely different genetic maps
- Segments are rearranged during evolution
- Each block represents a set of conserved genes

Dashed lines correspond to reversed segments

Genome Evolution

Suggested Models

1) *(slow shuffle)* Independent segmental duplication with rearrangement
2) *(big bang)* Whole genome duplication
Slow Shuffle

Ancestor A:

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0/

Inversion 1

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 6 5 4 3 2 1//0 9 8 8 9 0//1 2 3 4 5 6 7 8 9 0/

Inversion 2

/1 2 3 4 3 2 1//0 9 8 8 9 0//1 2 3 4 5 6 7 6 5 4 3 2 1//0 9 8 7 6 5 4 5 6 7 8 9 0/

Chromosomes of Descendant B:

- B1: /1 2 3 4 5 6 7 6 5 4 3 2 1/
- B2: /0 9 8 7 6 5 4 5 6 7 8 9 0/
- B3: /1 2 3 4 3 2 1/
- B4: /0 9 8 8 9 0/
Slow Shuffle

Ancestor A:

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0/

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0/

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0/

/1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0//1 2 3 4 5 6 7 8 9 0/

Chromosomes of Descendant C:

C1 /1 2 3 4 5 6 7 8 3 4 9 8 7 6 5 4 3 2 1/0 9 2 1/
C2 /0 9 8 7 6 5 4 3 2 1/
C3 /0 5 6 7 8 9 0/
C4 /0 9 2 1/
5.1 The Biological Problem

5.1.1 Modeling Conserved Synteny

When comparing genomes of two organisms such as mice and humans, we sometimes observe local similarities in the gene content and gene order. If two or more genes that occur together on a single chromosome in organism B also occur together on a particular chromosome in organism C, this occurrence is an example of **conserved synteny** (see Fig 1.4 in Chapter 1). As shown in Fig. 5.1, the gene orders may or may not be conserved within syntenic blocks that are distributed among different chromosomes in a related organism. We will use a simple example to show how such patterns can arise by genome rearrangement.

Figure 5.2 shows four different ancestral "chromosomes," each distinguished by a different color. Each ancestral chromosome is of identical size, and each has ten "genes" 0, 1, ..., 9. Each number/color combination represents a unique gene. This means that 2-green is not the same gene as 2-red. The telomeres are indicated by a single slash, `/`. We will model an evolutionary process indicated in the following diagram:

Ancestor A

Descendant B

Descendant C

inversion 1 inversion 2

inversion 3 inversion 4

To model the genome rearrangements, we line up the A chromosomes in tandem and then invert different portions of this combined string in two independent steps to obtain chromosomes of Descendant B. The same process is repeated by inverting different segments to produce Descendant C. By inversion, we mean that we take a portion of the sequence of genes and reverse their order relative to the surrounding genes. For example, if the middle three genes represented by 1, 2, 3 were inverted, the result would be 1, 4, 3, 2, 5.

Inversions of chromosomal regions are well-documented in most organisms for which there are genetic data. Fig. 5.2 (Following page) [This figure also appears in the color insert.] Model producing conserved synteny in descendant chromosomes B and C after independent inversion steps starting with ancestral chromosome A. Telomeres (/) divide the "genes" (numbered) into four chromosomes. Chromosomes are placed end-to-end before inversions (reversals) are performed to simulate translocation. Two inversion steps on the path leading to B and two inversion steps on the path leading to C are sufficient to generate syntenic relationships resembling those seen in Fig. 5.1.
Fig. 14.10. Identifying whole-genome duplication by between-genome comparison. The unduplicated genome is shown in panel A. Genes are represented as pentagons with a gradient of shading. After whole-genome duplication (panel B), two copies of every gene are produced (unshaded pentagons). Genes can mutate (panel C), and these mutations may destroy functions (indicated by broken bounding lines) or alter them (shaded genes labeled with primes). Deletions may occur (panel D), subject to the constraint that essential genes be retained. Alignment of chromosomes D1 and D2 with the unduplicated genome of another descendant of the shared common ancestor (panel E) supports the whole-genome duplication because nonparalogous genes in the two chromosomes appear interleaved and in the same order as genes in the unduplicated chromosome.
Fig. 14.10. Identifying whole-genome duplication by between-genome comparison. The unduplicated genome is shown in panel A. Genes are represented as pentagons with a gradient of shading. After whole-genome duplication (panel B), two copies of every gene are produced (unshaded pentagons). Genes can mutate (panel C), and these mutations may destroy functions (indicated by broken bounding lines) or alter them (shaded genes labeled with primes). Deletions may occur (panel D), subject to the constraint that essential genes be retained. Alignment of chromosomes D1 and D2 with the unduplicated genome of another descendant of the shared common ancestor (panel E) supports the whole-genome duplication because nonparalogous genes in the two chromosomes appear interleaved and in the same order as genes in the unduplicated chromosome.
Fig. 14.10. Identifying whole-genome duplication by between-genome comparison.

The unduplicated genome is shown in panel A. Genes are represented as pentagons with a gradient of shading. After whole-genome duplication (panel B), two copies of every gene are produced (unshaded pentagons). Genes can mutate (panel C), and these mutations may destroy functions (indicated by broken bounding lines) or alter them (shaded genes labeled with primes). Deletions may occur (panel D), subject to the constraint that essential genes be retained. Alignment of chromosomes D1 and D2 with the unduplicated genome of another descendant of the shared common ancestor (panel E) supports the whole-genome duplication because nonparalogous genes in the two chromosomes appear interleaved and in the same order as genes in the unduplicated chromosome.
Big Bang

Fig. 14.10. Identifying whole-genome duplication by between-genome comparison. The unduplicated genome is shown in panel A. Genes are represented as pentagons with a gradient of shading. After whole-genome duplication (panel B), two copies of every gene are produced (unshaded pentagons). Genes can mutate (panel C), and these mutations may destroy functions (indicated by broken bounding lines) or alter them (shaded genes labeled with primes). Deletions may occur (panel D), subject to the constraint that essential genes be retained. Alignment of chromosomes D1 and D2 with the unduplicated genome of another descendant of the shared common ancestor (panel E) supports the whole-genome duplication because nonparalogous genes in the two chromosomes appear interleaved and in the same order as genes in the unduplicated chromosome.

5.1.1 Modeling Conserved Synteny

When comparing genomes of two organisms such as mice and humans, we sometimes observe local similarities in the gene content and gene order. If two or more genes that occur together on a single chromosome in organism B also occur together on a particular chromosome in organism C, this occurrence is an example of conserved synteny (see Fig 1.4 in Chapter 1). As shown in Fig. 5.1, the gene orders may or may not be conserved within syntenic blocks that are distributed among different chromosomes in a related organism. We will use a simple example to show how such patterns can arise by genome rearrangement.

Figure 5.2 shows four different ancestral "chromosomes," each distinguished by a different color. Each ancestral chromosome is of identical size, and each has ten "genes" 0, 1, ..., 9. Each number/color combination represents a unique gene. This means that 2-green is not the same gene as 2-red. The telomeres are indicated by a single slash, /.

We will model an evolutionary process indicated in the following diagram:

Ancestor A

Descendant B

Descendant C

inversion 1

inversion 2

inversion 3

inversion 4

To model the genome rearrangements, we line up the A chromosomes in tandem and then invert different portions of this combined string in two independent steps to obtain chromosomes of Descendant B. The same process is repeated by inverting different segments to produce Descendant C. By inversion, we mean that we take a portion of the sequence of genes and reverse their order relative to the surrounding genes. For example, if the middle three genes represented by 12345 were inverted, the result would be 14325.

Inversions of chromosomal regions are well-documented in most organisms for which there are genetic data.

Fig. 5.2 (Following page) [This figure also appears in the color insert.] Model producing conserved synteny in descendant chromosomes B and C after independent inversion steps starting with ancestral chromosome A. Telomeres (/) divide the "genes" (numbered) into four chromosomes. Chromosomes are placed end-to-end before inversions (reversals) are performed to simulate translocation. Two inversion steps on the path leading to B and two inversion steps on the path leading to C are sufficient to generate syntenic relationships resembling those seen in Fig. 5.1.
Genome Rearrangement: Sorting by Reversals
Reversal
Human X vs Mouse X

![Genome Diagram](image)

Fig. 5.3. Synteny blocks shared by human and mouse X chromosomes. The arrowhead for each block indicates the direction of increasing coordinate values for the human X chromosome. Reprinted, with permission, from Pevzner P and Tesler G (2003) Genome Research 13:37–45. Copyright 2003 Cold Spring Harbor Laboratory Press.

We now transform genome $y$ into genome $y_4$ through a series of four reversal steps of our choice, saving each intermediate genome $y_1$, $y_2$, and $y_3$. $y_2$ is produced by applying a second reversal to $y_1$, which resulted from the first reversal, and so on.

```r
y1 <- y
y2 <- y1
```
Mitch. Cabbage vs Turnip

Cabbage

Turnip
Mitch. Cabbage vs Turnip

Cabbage

Turnip

Sunday, February 23, 14
Mitch. Cabbage vs Turnip

Cabbage

Turnip
Mitch. Cabbage vs Turnip

Cabbage

Turnip

Reversal distance = 3
> Nearly all of the small human ch20 corresponds to a portion of the large mouse ch2.
> Portions corresponding to Mouse ch2 are also found on Human chromosome 2, 9, 10, 11, and 15

How to find distance?

There is a package

- Masking repeats
- Finding CpG islands
- Finding genes
- Genome Alignment
- .....