Ricerca di blocchi di sintenia

- Raglianti Marco [ragliant@cli.di.unipi.it]
- Santoro Roberto [santoror@cli.di.unipi.it]

Cosa vedremo

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

Cosa vedremo

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

Definizioni

- Omologia: due geni si dicono omologhi se hanno sequenza simile come risultato della derivazione dallo stesso gene ancestrale
  - Ortologia: omologia derivante da un evento di speciazione. Le copie di geni divergenti, nelle specie risultanti, sono dette ortologhe
  - Paralogia: omologia derivante da un evento di duplicazione genica.

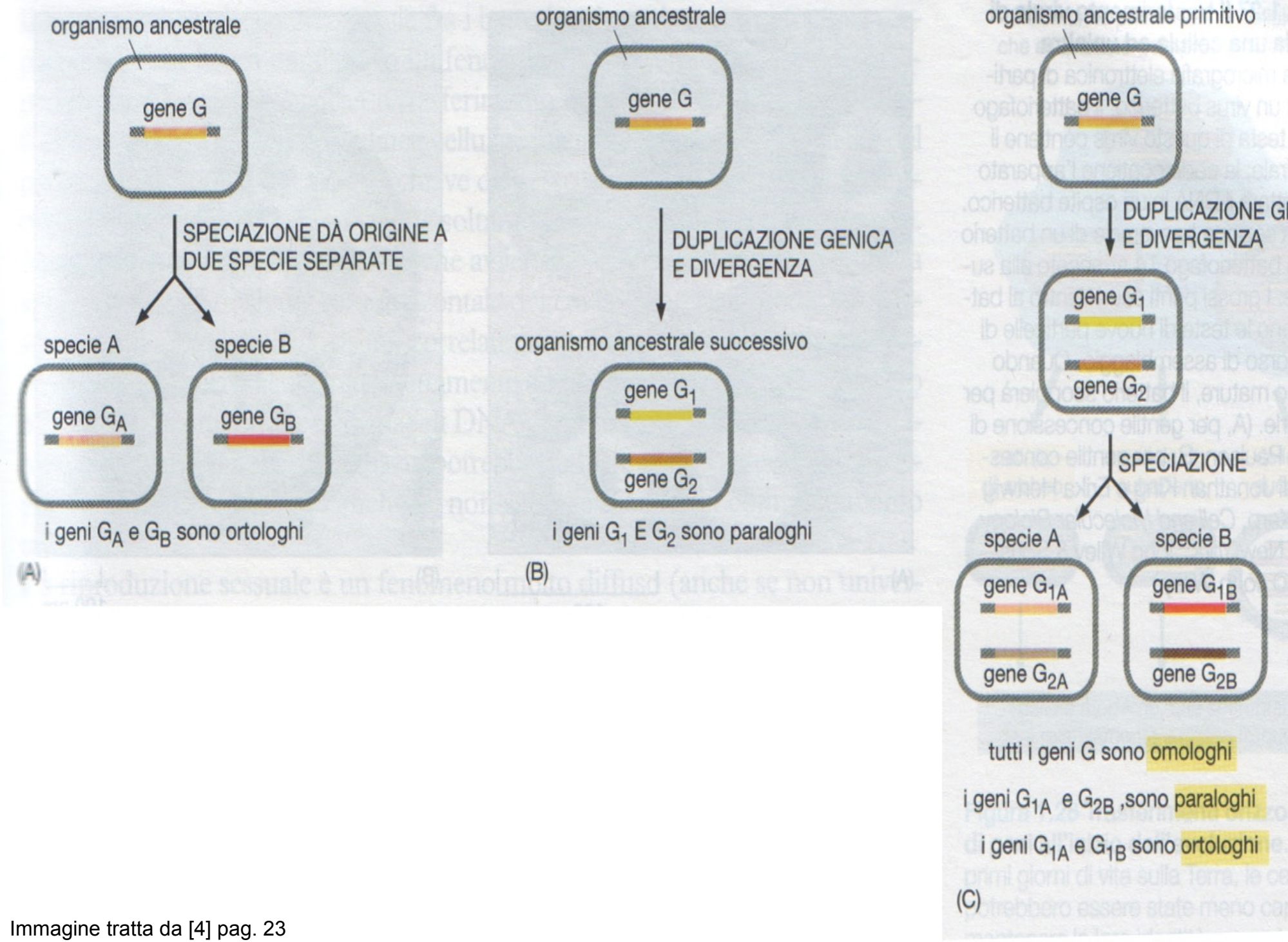


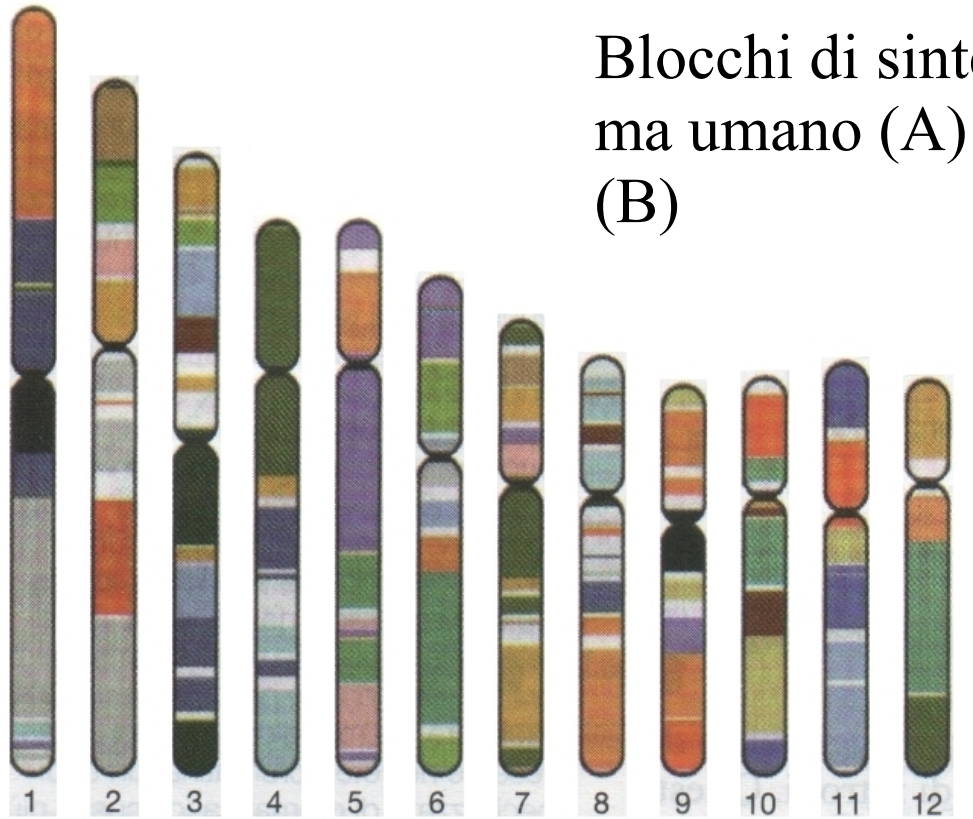
Immagine tratta da [4] pag. 23

- Sequenze simili hanno spesso funzioni simili (o anche identiche)
- Sequenze ortologhe sono necessariamente in specie differenti
- Nelle sequenze paraloghe, a causa della mancanza di *pressione selettiva* su una delle copie del gene, questa è libera di mutare e acquisire nuove funzioni

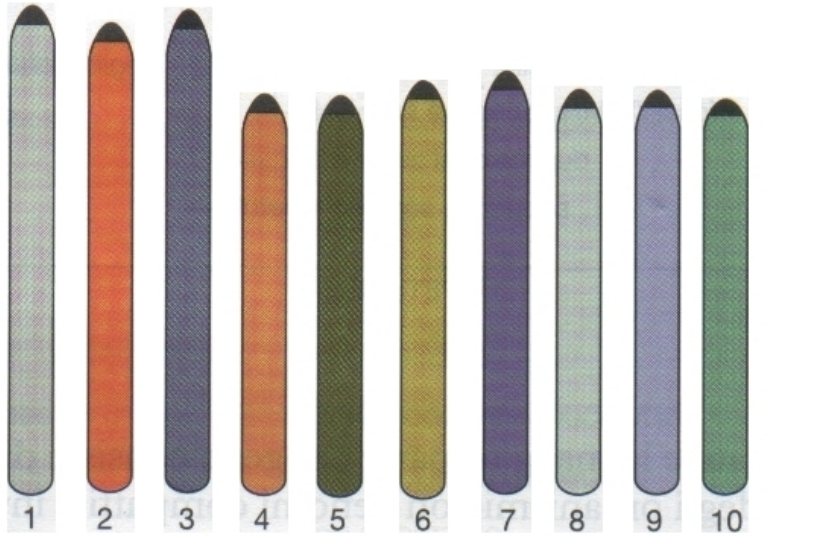
Marker e blocchi di sintenia

- Marker: sequenza di DNA
  - Nota
  - Facilmente identificabile
  - Associata ad una parte di genoma
- Blocchi di sintenia: regioni del genoma dove l'ordine dei geni viene conservato, come risultato della discendenza da un antenato comune

Blocchi di sintenia tra il genoma umano (A) e quello di topo (B)



(A)



(B)

Immagine tratta da [4] pag. 216

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

Aspetti biologici

- Rottura della sintenia (Synteny disruption)
  - Errori di mapping
  - Distanza evoluzionistica
- Alto numero di marker consecutivi in comune indica, con maggiore probabilità, sintenia.
- Basso numero può indicare elevata frammentazione dovuta ad elevato riarrangiamento genomico
  - Caso limite: un solo marker (errore di mapping?)

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

Dati del problema

- Input: Due genomi di specie differenti rispettivamente con χ_1 e χ_2 cromosomi
  - n_1 ed n_2 numero totale di marker
  - n numero di marker distinti in comune
    - $n_1 \geq n$ e $n_2 \geq n$
- I marker sono numerati univocamente sul primo genoma (etichettando gli insiemi di paralogia)
- Marker sul secondo genoma hanno la numerazione dei rispettivi omologhi nel primo

Genoma 1

c.1	abcdef	$\chi_1 = 3$
c.2	lmnoprq	$n_1 = 18$
c.3	wdxyz	

Genoma 2

c.1	lbcdpz	$\chi_2 = 4$
c.2	-x-q-o-mbc	$n_2 = 20$
c.3	wde-fry	
c.4	na	

Definizioni

- Strip: sequenza di h ≥ 2 marker consecutivi e contigui sia su un cromosoma del primo che su un cromosoma del secondo genoma
  - Definita rispetto allo stato corrente dei due genomi (l'algoritmo ridurrà le dimensioni dei due genomi)
- Caso particolare: marker in insiemi di paralogia
  - Cromosoma su genoma 1: xlabcdo
  - Cromosoma su genoma 2: rgabc'c”df
  - abcd è una strip (ricorda che c è etichettato come insieme di paralogia)

Definizioni

- Pre-strip: sequenza di $h \geq 2$ marker consecutivi (ma non necessariamente contigui) completa
  - Una pre-strip è completa sse non esiste alcun marker (di orientamento appropriato) su entrambi i cromosomi, tra due marker consecutivi
- Pure strip: pre-strip che è una strip nel genoma originale e non è contenuta in alcuna altra strip
- Pre-strip e pure strip sono definite unicamente rispetto al genoma originale
- Forward e reverse order

Originali

Genoma 1

c.1	abcdef
c.2	lmnoprq
c.3	wdxyz

Genoma 2

c.1	lbcdpz
c.2	-x-q-o-mbc
c.3	wde-fry
c.4	na

Pre-strip:
bcd, bc, cd,
moq, mo, oq,
wdy, wd, dy,
lp, de, dz

(Pure) strip:
bcd, bc, de, wd

Sottosequenze comuni non pre-strip:
bd, mq, wy

Problema MSR

Maximal Strip Recovery (MSR)

Dati due genomi, scartare alcuni sottoinsiemi di marker, lasciando soltanto marker in strip disgiunte $S_1, \ldots, S_r$ di lunghezze $h_1, \ldots, h_r$ tali che, nei genomi così ridotti,

$$\sum_{i=1}^{r} h_i$$ sia massimizzata

- $h_1, \ldots, h_r$ stime delle lunghezze dei blocchi di sintenia che contengono le strip

- equivale a costruire un insieme di stringhe compatibili contenenti la maggior informazione possibile

L'algoritmo (Overview)

- Generazione di un sottoinsieme significativo di pre-strip (in tempo polinomiale)
- Costruzione del grafo di compatibilità (o del complementare grafo di conflitto)
- Risoluzione del problema Maximum Weight Clique (o del Maximum Weight Independent Set)
  - La soluzione fornisce l'insieme ottimale di pre-strip (ossia la soluzione del problema MSR)
- Ripristino di marker compatibili

- Proposizione: Ogni pre-strip P può essere rappresentata con una sequenza di termini nella forma p, 11, 1p, p1, 111, 1p1
  - “p” rappresenta una pure strip
  - “1” rappresenta un marker (singleton) non in una pure strip in P
- Generiamo tutte le pre-strip nella forma p, 11, 1p, p1, 111, 1p1
  - La soluzione del problema MSR su tali pre-strip è ***equivalente*** a quella ottenuta a partire dai genomi originali
  - Tutti i marker non appartenenti a tali pre-strip non verranno più considerati ai fini della soluzione del problema MSR

Originali

Genoma 1

c.1	abcdef
c.2	lmnoprq
c.3	wdxyz

Genoma 2

c.1	lbcdpz
c.2	-x-q-o-mbc
c.3	wde-fry
c.4	na

Pre-strip:
bcd, bc, cd,
moq, mo, oq,
wdy, wd, dy,
lp, de, dz

(Pure) strip:
bcd, bc, de, wd

Sottosequenze comuni non pre-strip:
bd, mq, wy

Ridotti

Genoma 1

abcd
lmoq
wdyz

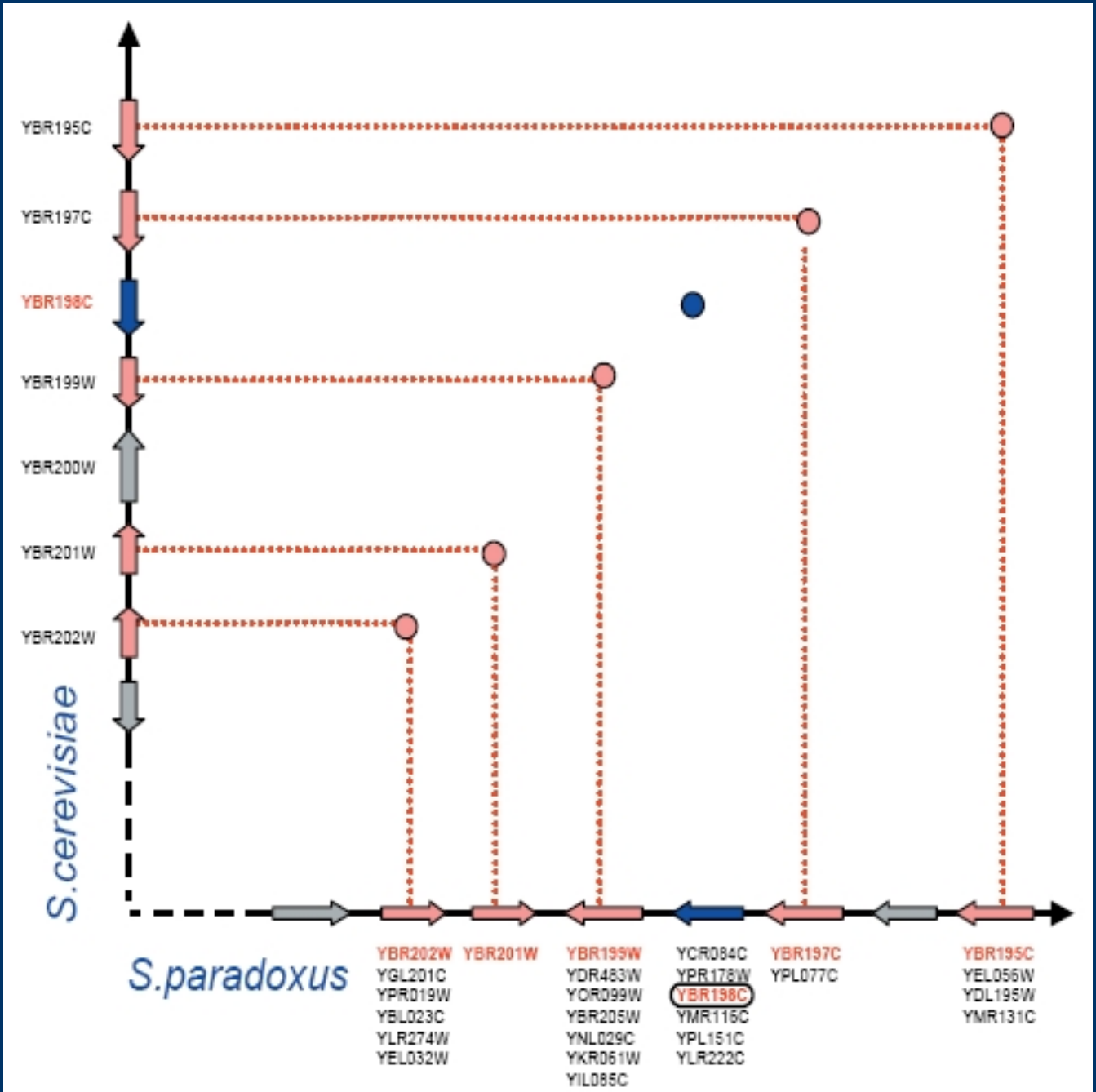
Genoma 2

lbcdz
-q-o-m
wdy
a

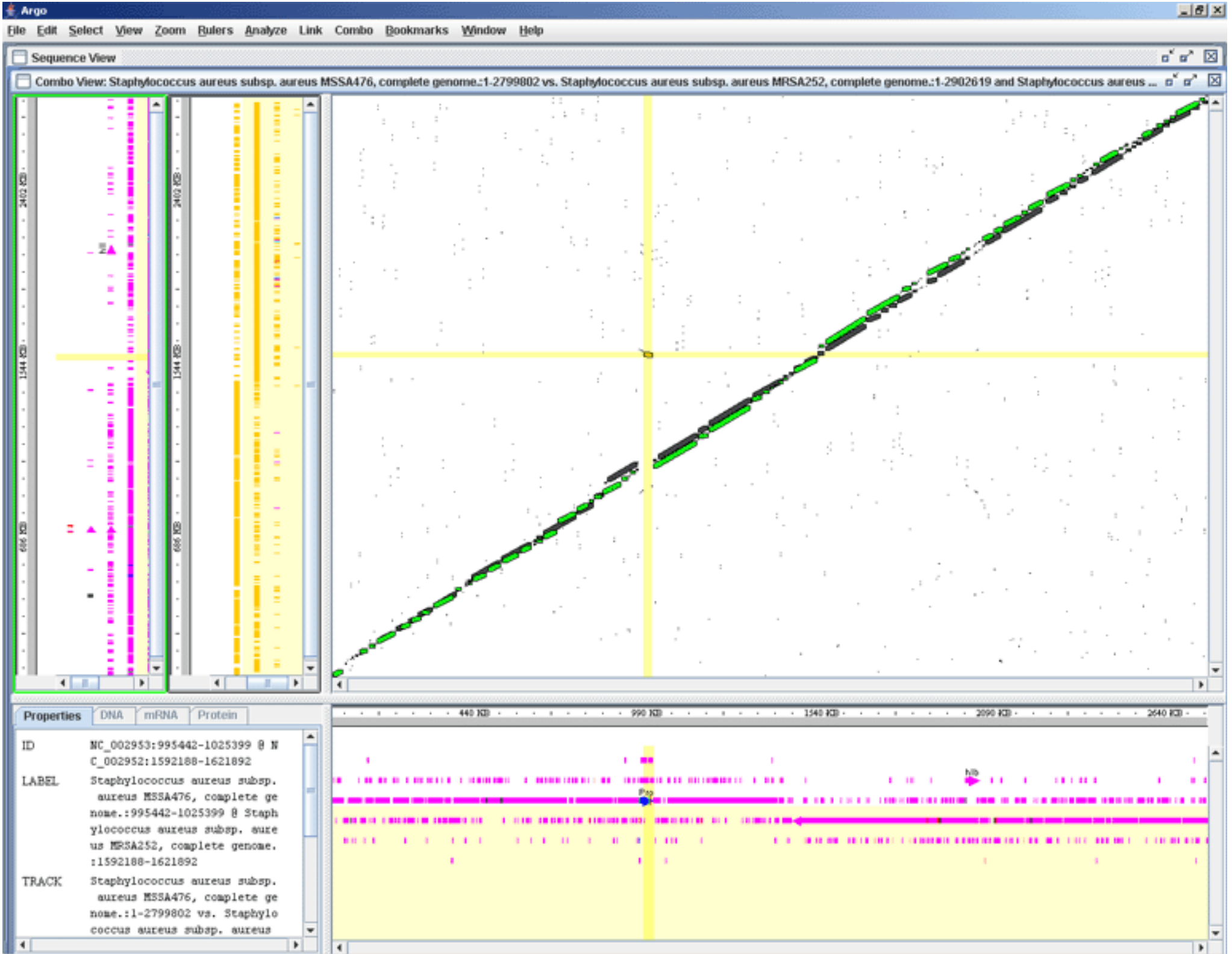
Strip:
bcd, moq, wdy

Singleton non in pre-strip ma compatibili:
a, l, z

Scartati come rumore:
e, f, n, p, r, x

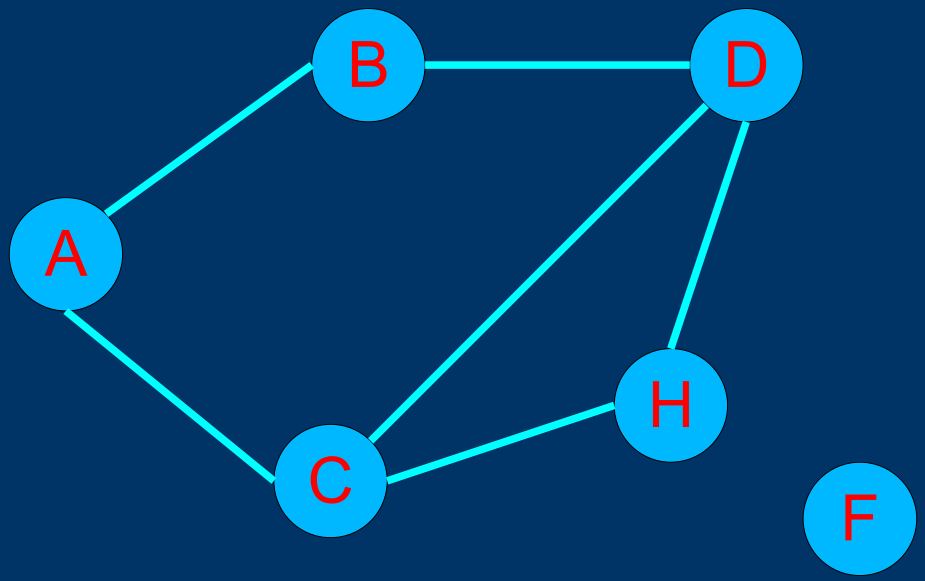


- L'algoritmo individua inizialmente le pure strip e i singleton
- A partire da questi genera le pre-strip della forma p, 11, 1p, p1, 111, 1p1


Argo
File Edit Select View Zoom Rulers Analyze Link Combo Bookmarks Window Help
Sequence View
Combo View: Staphylococcus aureus subsp. aureus MSSA476, complete genome.:1-2799802 vs. Staphylococcus aureus subsp. aureus MRSA252, complete genome.:1-2902619 and Staphylococcus aureus ...
2402 KB
1344 KB
686 KB
Properties
DNA
mRNA
Protein
ID NC_002953:995442-1025399 @ NC_002952:1592188-1621892
LABEL Staphylococcus aureus subsp. aureus MSSA476, complete genome.:995442-1025399 @ Staphylococcus aureus subsp. aureus MRSA252, complete genome.:1592188-1621892
TRACK Staphylococcus aureus subsp. aureus MSSA476, complete genome.:1-2799802 vs. Staphylococcus aureus subsp. aureus
440 KB
990 KB
1540 KB
2090 KB
2640 KB
hlb
Pig

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

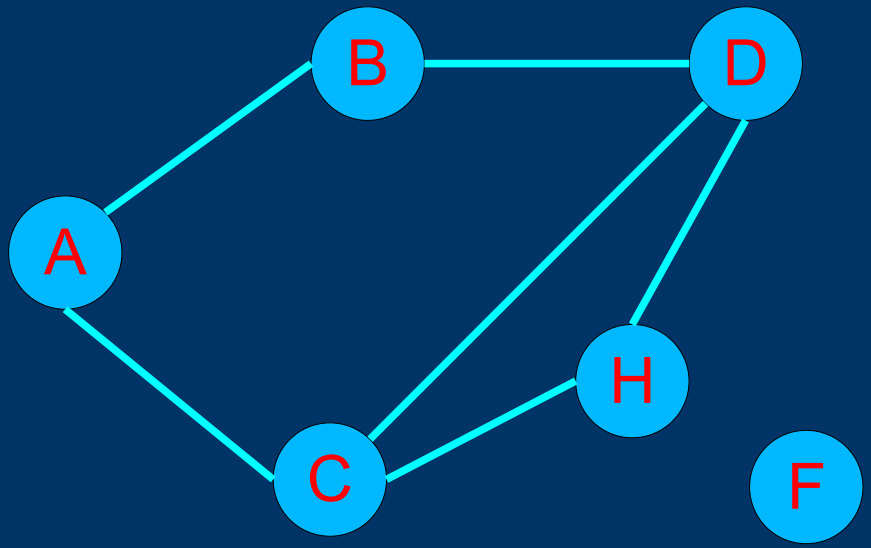
- Un grafo è un insieme di elementi detti nodi (o vertici) collegati fra loro da archi
  - Più formalmente, si dice grafo una coppia ordinata G = (V, E) di insiemi, con V insieme dei nodi ed E insieme degli archi, tali che gli elementi di E siano coppie di elementi di V



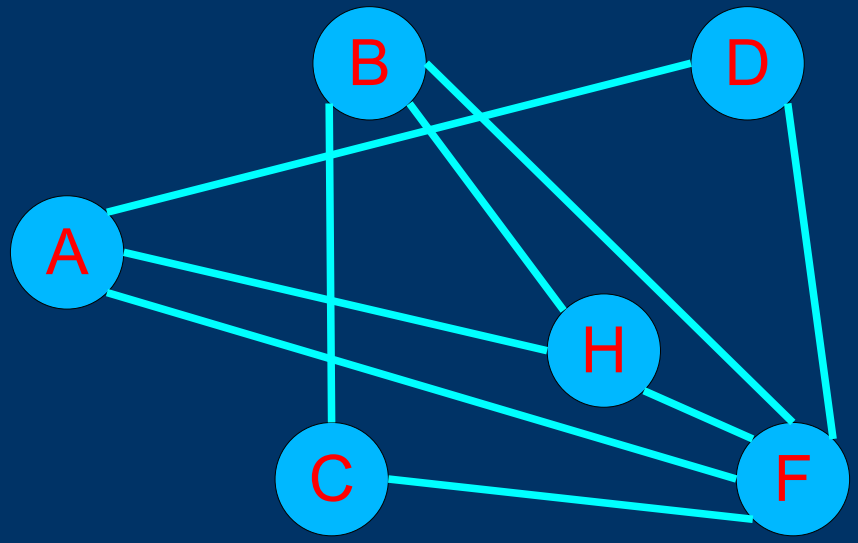
G = (V, E)
V = { A, B, C, D, H, F }
E = { (A,B), (A,C), (B,D), (D,C), (D,H), (C,H) }

- Dato un grafo G = (V, E), il grafo G' = (V, E') è il suo complementare se e solo se:
  - $E' = \{(u,v) : u \neq v \in V , (u,v) \notin E\}$

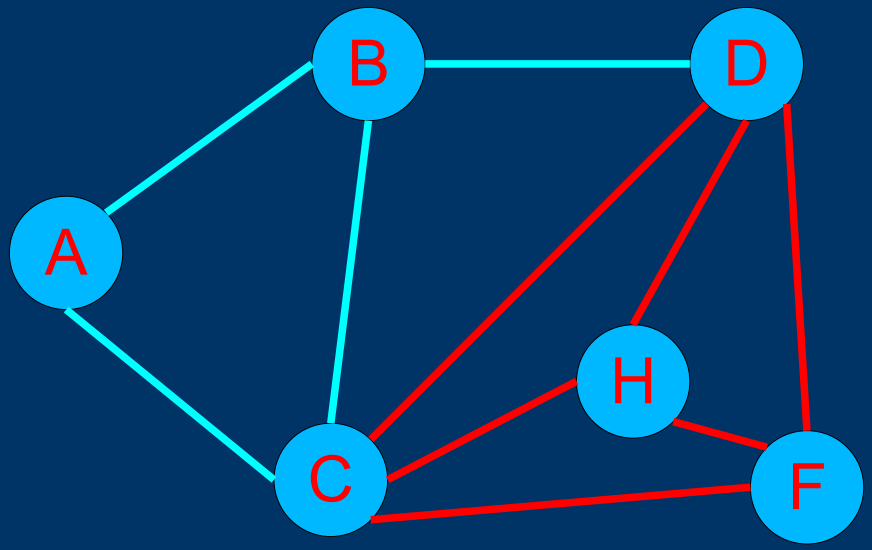
Grafo

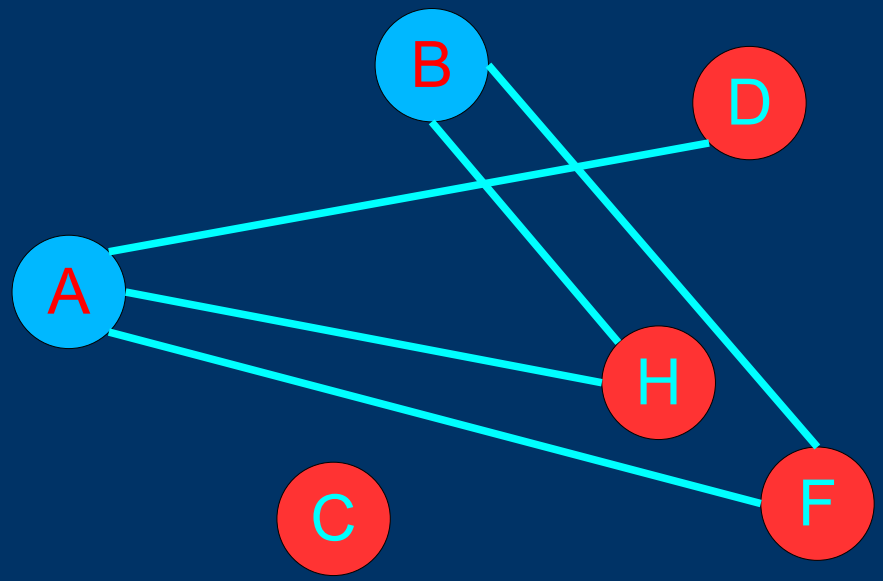


Grafo Complementare



- Una cricca (clique) in un grafo G è un insieme V di vertici tale che, per ogni coppia di vertici in V, esiste un arco che li collega
- Concetto complementare: insieme indipendente (ad ogni clique corrisponde un insieme indipendente nel grafo complementare)
- Un insieme indipendente (independent set) in un grafo G è un insieme V di vertici tale che, per ogni coppia di vertici in V, non esiste alcun arco che li collega





- Due pre-strip P e Q si dicono incompatibili se:
  - P e Q hanno almeno un marker in comune, oppure
  - P contiene un marker fra due marker successivi in Q, in uno dei due genomi
    - Es. Genoma 1:

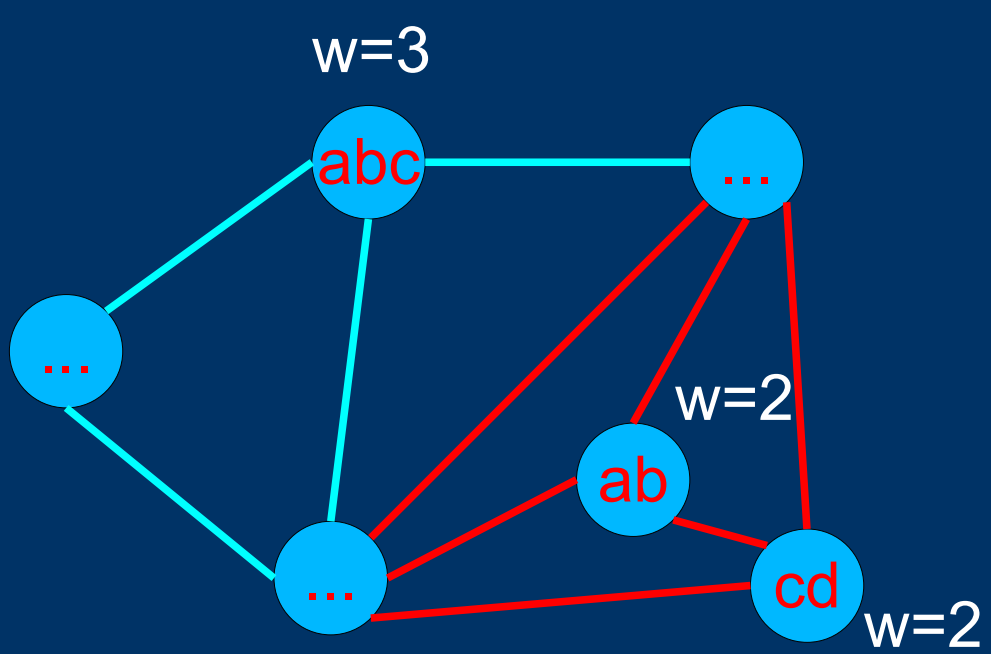
      Cromosoma 1: ...abc...

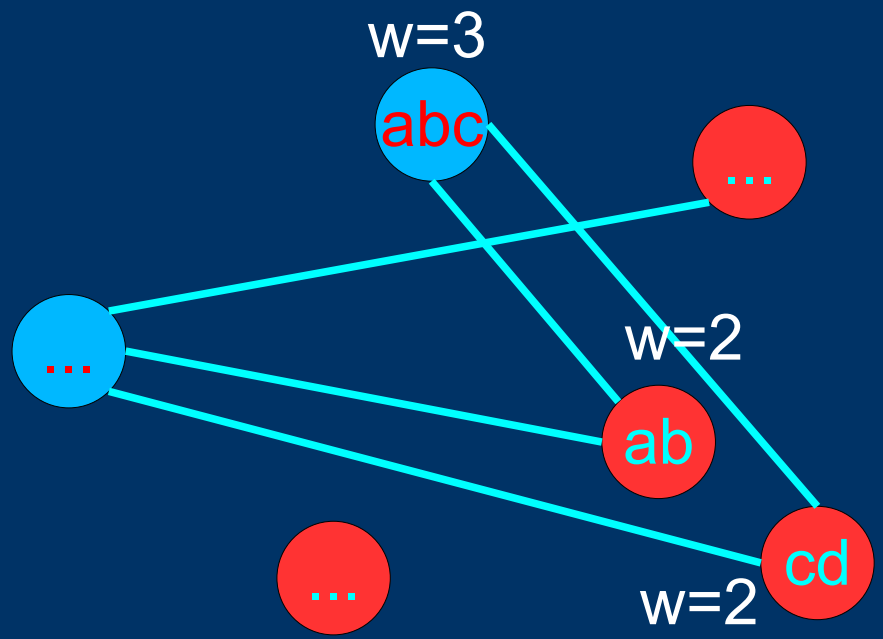
      Cromosoma 2: ...xyz...

      Genoma 2:

      Cromosoma 1: ...abxycz...
- Grafo di compatibilità:
  - Un nodo per ogni pre-strip
  - Peso di un nodo pari al numero di marker nella pre-strip
  - Pre-strip compatibili collegate da un arco

- Grafo di conflitto:
  - pre-strip incompatibili collegate da un arco
- È il complemento del grafo di compatibilità





Problema MWC

Maximum Weight Clique (MWC)

Dato un grafo con pesi positivi associati ai vertici, trovare una clique tale che la somma dei pesi dei vertici appartenenti alla clique sia massima

- La soluzione del problema MWC sul grafo di compatibilità ***equivale*** alla soluzione del problema MSR

Problema MWIS

Maximum Weight Independent Set (MWIS)

Dato un grafo con pesi positivi associati ai vertici, trovare un independent set tale che la somma dei pesi dei vertici appartenenti ad esso sia massima

- La soluzione del problema MWIS sul grafo di conflitto ***equivale*** quindi alla soluzione del problema MSR
- Equivalente al problema MWC sul grafo di compatibilità

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

- La soluzione di MSR è incompatibile con le pre-strip non in essa, ma non necessariamente con:
  - alcune parti di tali pre-strip; oppure
  - i marker che non compaiono in alcuna pre-strip (scartati inizialmente)
- I marker compatibili sono individuati mediante tecniche di genome rearrangement analysis
  - Aumento della distanza genomica in caso di incompatibilità

Originali

Genoma 1
c.1 abcde
c.2 fghklm

Genoma 2
c.1 ab-l-k
c.2 fgh-mcde

Strip restituite da MWC:
ab, cde, fgh, kl

Output

Output Genoma 1
ab cde
fgh kl

Output Genoma 2
ab -l-k
fgh cde

Distanza = 2
(Sia prima che dopo il ripristino del marker ***m***)

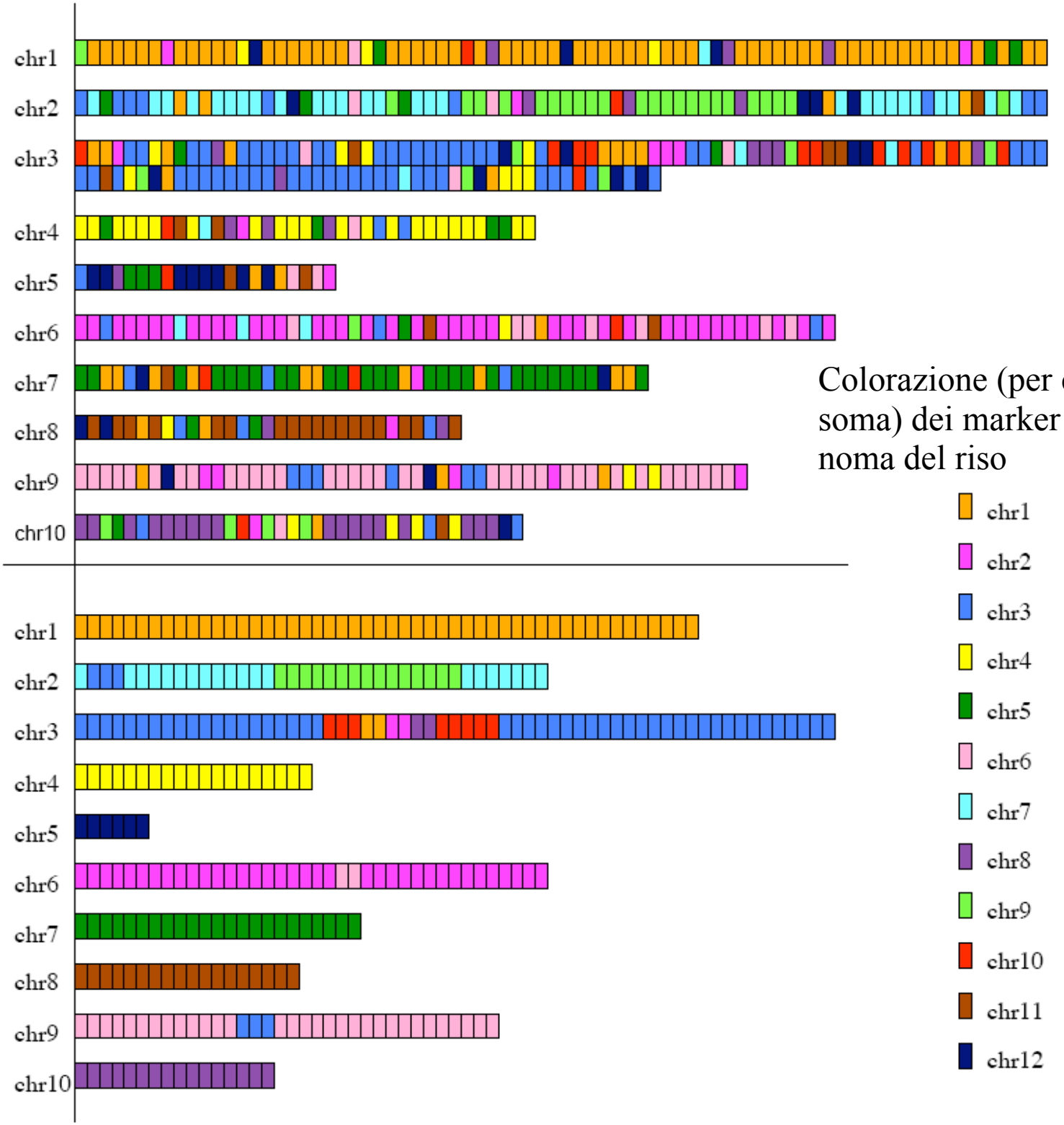
1 inversione
1 traslocazione

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)

- Il numero di pre-strip è esponenziale nel numero di marker nei genomi
- Il numero di pre-strip della forma p, 11, 1p, p1, 111, 1p1 invece cresce come una funzione polinomiale nel numero di marker nei genomi
- L'algoritmo genera tali pre-strip in tempo di $O(n^4)$ (alcasopessimo)

- Il collo di bottiglia dell'algoritmo consiste nella soluzione del problema MWC (MWIS)
  - NP-Hard
  - Risoluzione mediante euristiche
  - La riduzione del numero di pre-strip date in input velocizza l'esecuzione
    - È possibile ridurre ulteriormente il sottoinsieme di pre-strip in base ad alcune considerazioni sulla compatibilità fra di esse

- Concetti di base (e definizioni)
- Aspetti biologici
- L'algoritmo
  - Presentazione
  - Due approcci complementari
  - Ripristino dei marker
  - Complessità (e punti critici)
- Applicazione sperimentale (un esempio)



Genoma del sorgo

Colorazione (per cromosoma) dei marker sul genoma del riso

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12

Blocchi di sintenia (tra sorgo e riso)

constraints	pre-strips	running time	strips output	markers in output	markers restored	total markers	distance
$G < 5$ no $11's$	441	2 minutes	97	286	17	303	49
$G < 4$	543	24 hours	124	309	13	322	69
$G < 3$	428	29 minutes	123	302	13	315	63
$G < 2$	257	6 seconds	115	282	36	318	65

- Input:
  - 481 marker del riso
  - 567 marker del sorgo
- 481 marker distinti
- Occorrono comunque vincoli sulla dimensione dei gap per rendere trattabile il problema

constraints	pre-strips	running time	strips output	markers in output	markers restored	total markers	distance
$G < 5$ no $11's$	441	2 minutes	97	286	17	303	49
$G < 4$	543	24 hours	124	309	13	322	69
$G < 3$	428	29 minutes	123	302	13	315	63
$G < 2$	257	6 seconds	115	282	36	318	65

- Al diminuire della dimensione del gap:
  - diminuisce il tempo di esecuzione dell'algoritmo
  - Non cambia il numero di strip e marker in output dopo il ripristino

constraints	pre-strips	running time	strips output	markers in output	markers restored	total markers	distance
$G < 5$ no $11's$	441	2 minutes	97	286	17	303	49
$G < 4$	543	24 hours	124	309	13	322	69
$G < 3$	428	29 minutes	123	302	13	315	63
$G < 2$	257	6 seconds	115	282	36	318	65

- Eliminando pre-strip della forma 11 (la prova più debole di sintenia dopo il singleton)
  - diminuisce il numero di pre-strip
  - diminuisce il tempo di esecuzione

Bibliografia

- **[1] Models in comparative genomics: genome correspondence, gene identification and motif discovery** di M.Kellis and N.Patterson and B.Birren and B.Berger and E.S.Lander, Journal of Computational Biology 11(2) 319-355, 2004
- **[2] Removing Noise and Ambiguities from Comparative Maps in Rearrangement Analysis** di C.Zheng and Q.Zhu and D.Sankoff, to appear in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007
- **[3] Algorithms for the extraction of Synteny Blocks from Comparative Maps** di V.Choi and C.Zheng and Q.Zhu and D.Sankoff, 7th Workshop on Algorithms in Bioinformatics (WABI), 277-288, 2007
- **[4] Biologia Molecolare della cellula** di Alberts, Johnson, Lewis, Raff, Roberts, Walter (Zanichelli)
- **[5] Algorithmic Aspects of Bioinformatics** di H.J.Böckenhauer, D.Bongartz (Springer, 2007)