# Optimal *k*-mer superstrings for peptide identification from tandem mass spectra

**Nathan Edwards**

**Center for Bioinformatics and Computational Biology**

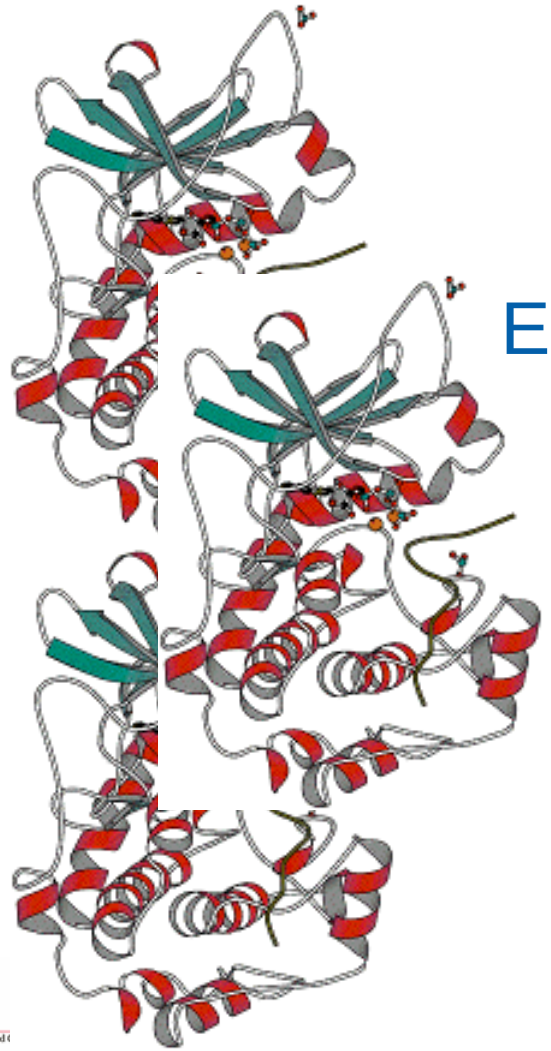
UMIACS
University of Maryland Institute for Advanced Computer Studies

UNIVERSITY OF MARYLAND 18 56

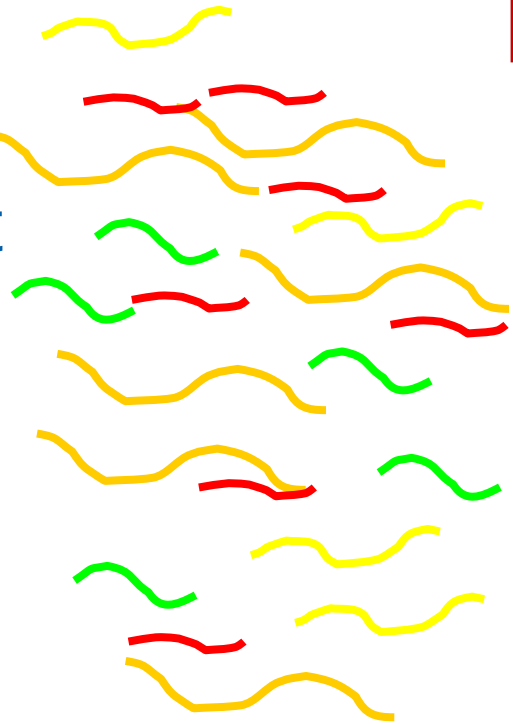
# Proteomics

- Study of observed proteins
  - How much of each?
  - What protein is it?
- Usually a combination of
  - Wet-lab biological sample manipulation
  - Mass spectrometry
  - Data analysis

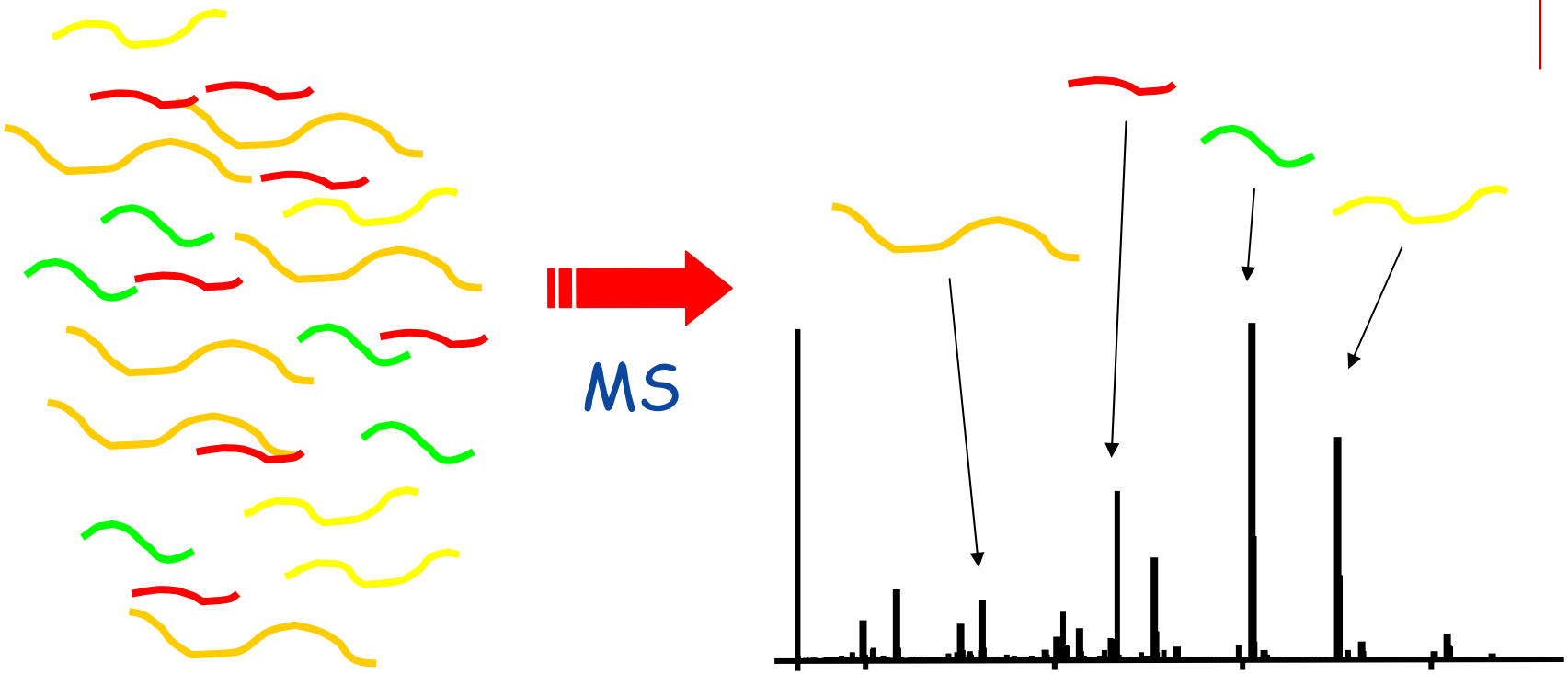
# Sample Preparation



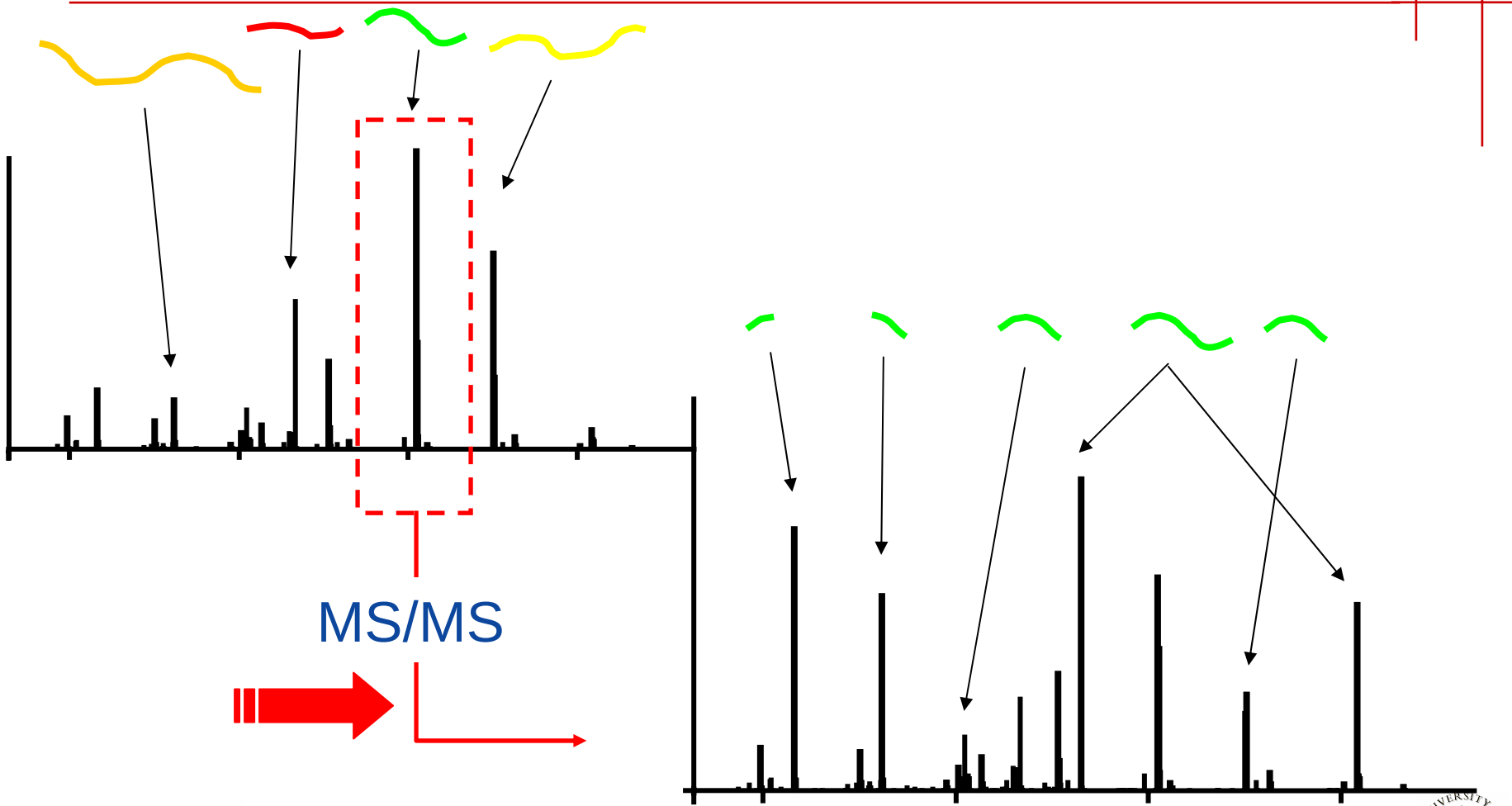
Enzymatic Digest
and
Fractionation



# (Single Stage) Mass Spectrometry



# Tandem Mass Spectrometry

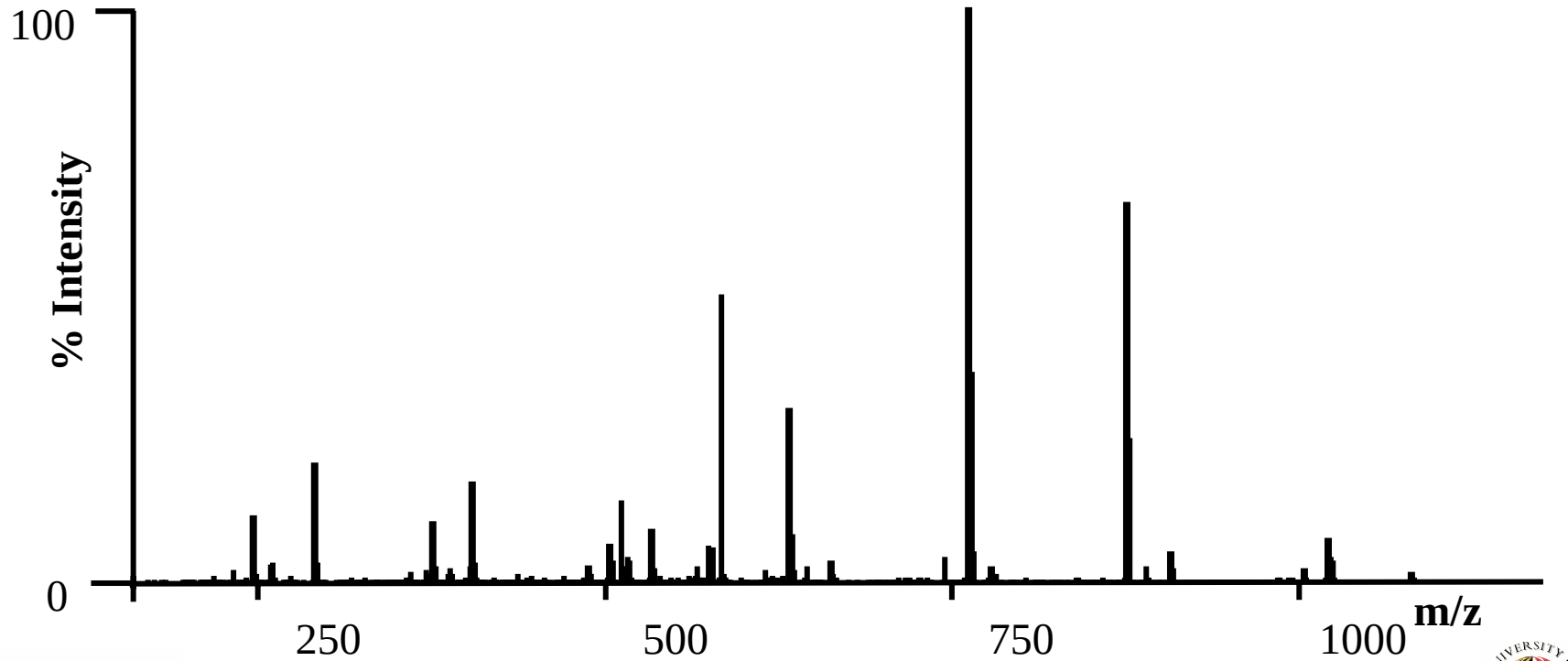


# Peptide Fragmentation

Peptide: S-G-F-L-E-E-D-E-L-K

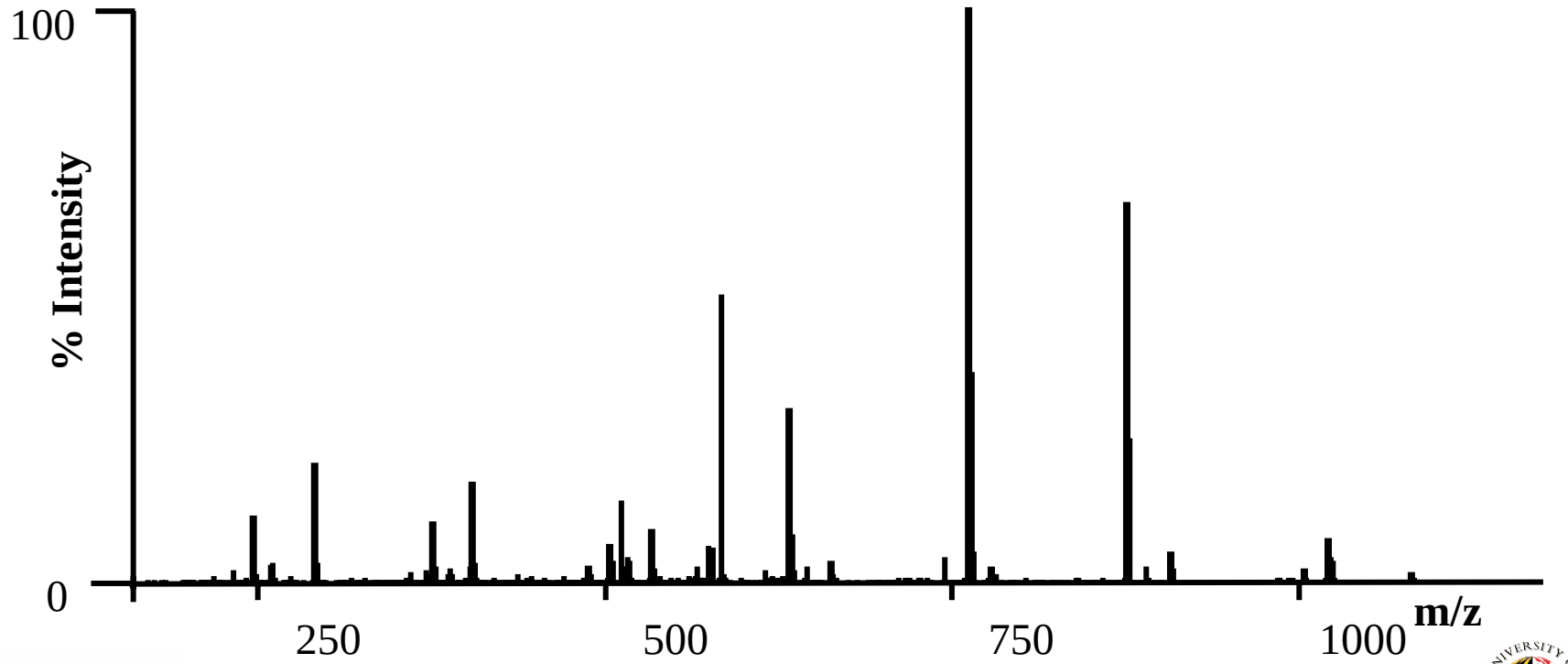
| MW | ion | | | ion | MW |
|---|---|---|---|---|---|
| 88 | $b_1$ | S | GFLEEDELK | $y_9$ | 1080 |
| 145 | $b_2$ | SG | FLEEDELK | $y_8$ | 1022 |
| 292 | $b_3$ | SGF | LEEDELK | $y_7$ | 875 |
| 405 | $b_4$ | SGFL | EEDELK | $y_6$ | 762 |
| 534 | $b_5$ | SGFLE | EDELK | $y_5$ | 633 |
| 663 | $b_6$ | SGFLEE | DELK | $y_4$ | 504 |
| 778 | $b_7$ | SGFLEED | ELK | $y_3$ | 389 |
| 907 | $b_8$ | SGFLEEDE | LK | $y_2$ | 260 |
| 1020 | $b_9$ | SGFLEEDEL | K | $y_1$ | 147 |

# Peptide Fragmentation



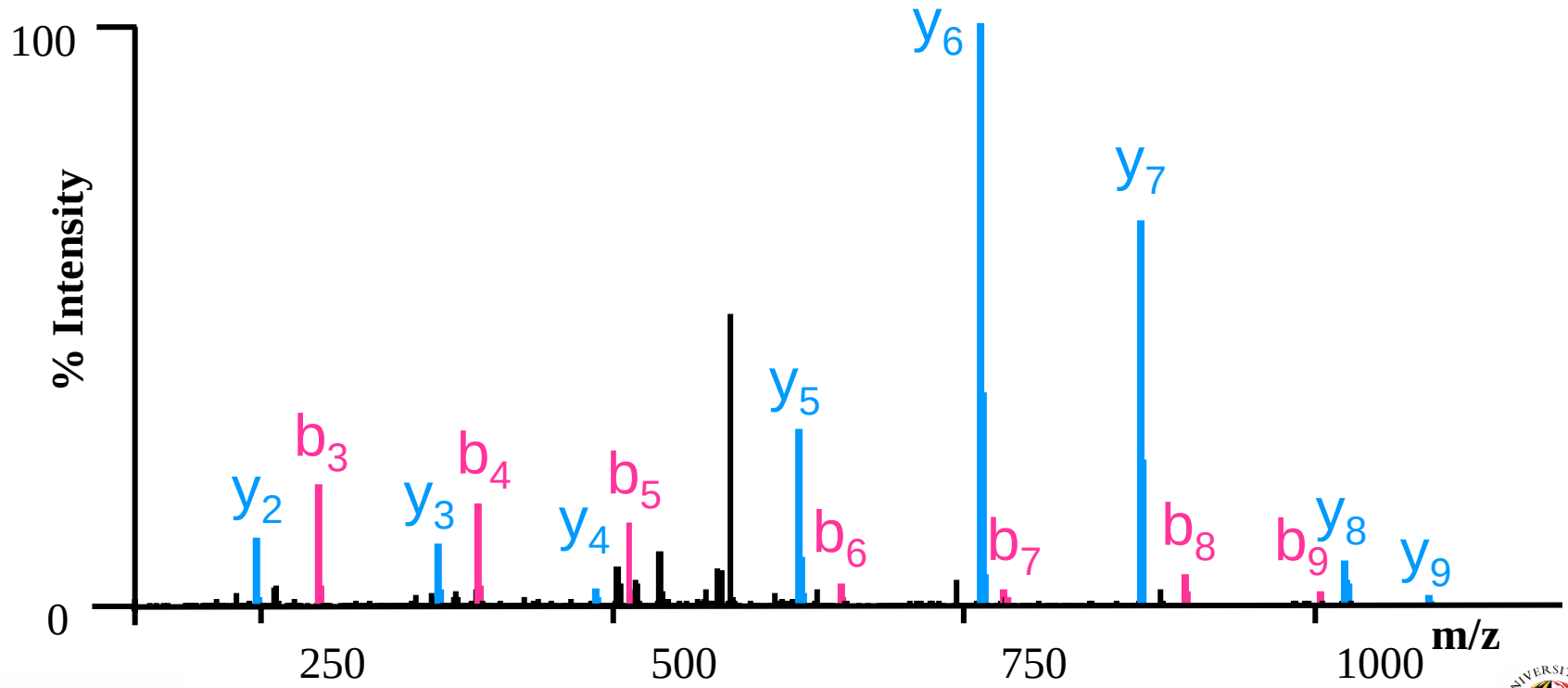
# Peptide Fragmentation

| 88 | 145 | 292 | 405 | 534 | 663 | 778 | 907 | 1020 | 1166 | b ions |
|---|---|---|---|---|---|---|---|---|---|---|
| S | G | F | L | E | E | D | E | L | K | |
| 1166 | 1080 | 1022 | 875 | 762 | 633 | 504 | 389 | 260 | 147 | y ions |



# Peptide Fragmentation

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 88 | 145 | 292 | 405 | 534 | 663 | 778 | 907 | 1020 | 1166 | b ions |
| S | G | F | L | E | E | D | E | L | K | |
| 1166 | 1080 | 1022 | 875 | 762 | 633 | 504 | 389 | 260 | 147 | y ions |



# Peptide Identification

Given:

- The mass of the parent ion
- The MS/MS spectrum

Output:

- The amino-acid sequence of the peptide

# Sequence Database Search

- Compares peptides from a protein sequence database with spectra
- Filter peptide candidates by
  - Parent mass
  - Digest motif
- Score each peptide against spectrum
  - Generate all possible peptide fragments
  - Match putative fragments with peaks
  - Score and rank

# Peptide Candidates

- Parent ion
  - Typically < 3000 Da
- Tryptic Peptides
  - Cut at K or R
- Search engines
  - Don’t handle > 4+ well
  - Long peptides don’t fragment well
- # of distinct 30-mers ($N_{30}$) upper bounds total peptide content

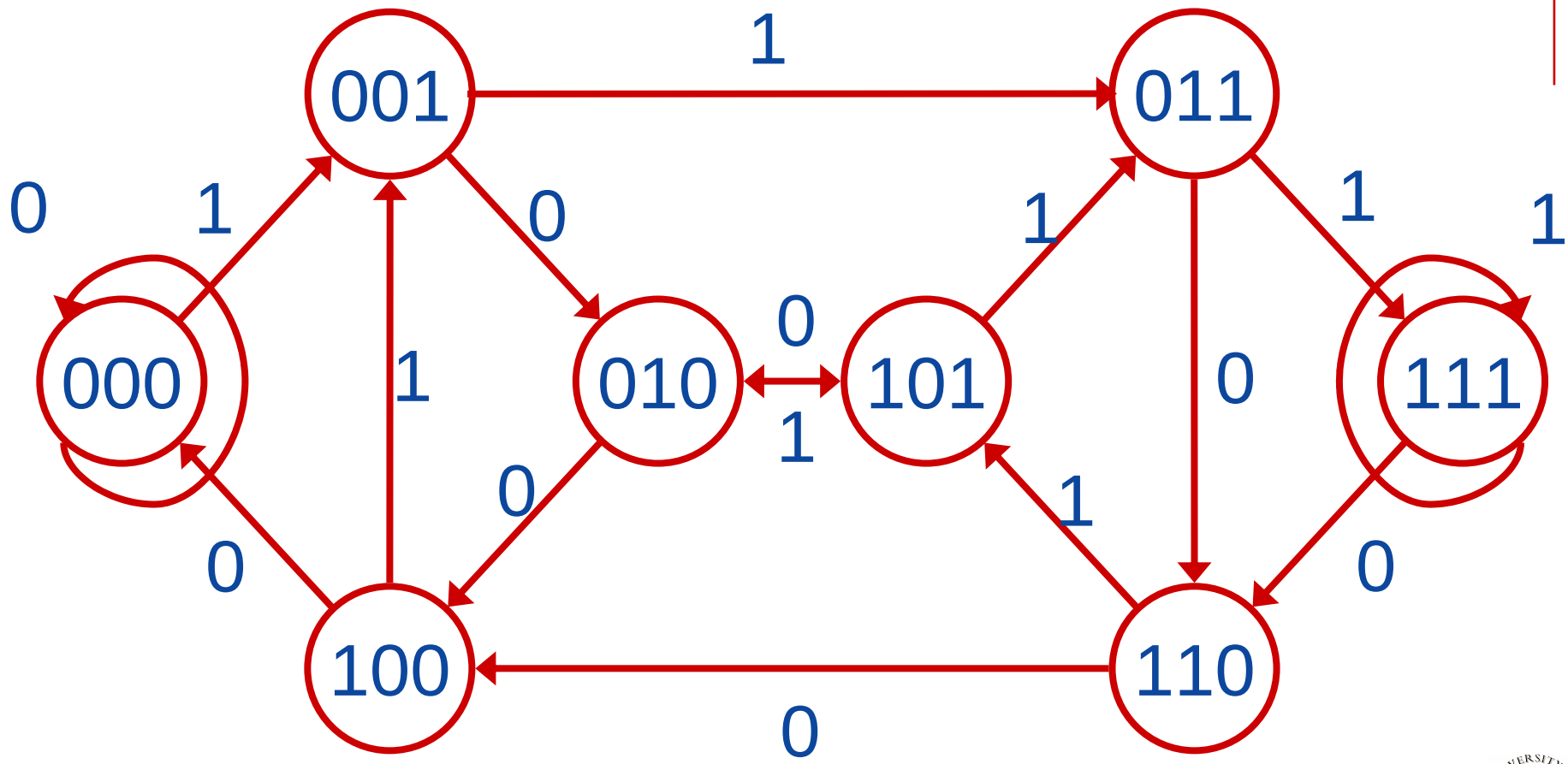
# de Bruijn Sequences

de Bruijn sequences represent all words of length $k$ from some alphabet $A$.

$A = \{0,1\}$, $k = 3$: $s = 0001110100$

$A = \{0,1\}$, $k = 4$: $s = 0000111101011001000$

# de Bruijn Graph: $A = \{0,1\}$, $k = 4$



# de Bruijn Sequences & Graphs

de Bruijn graphs ($k$,$A$):

- Edges represent length $k$ words from $A$
- Each node has
  - in degree $|A|$
  - out degree $|A|$

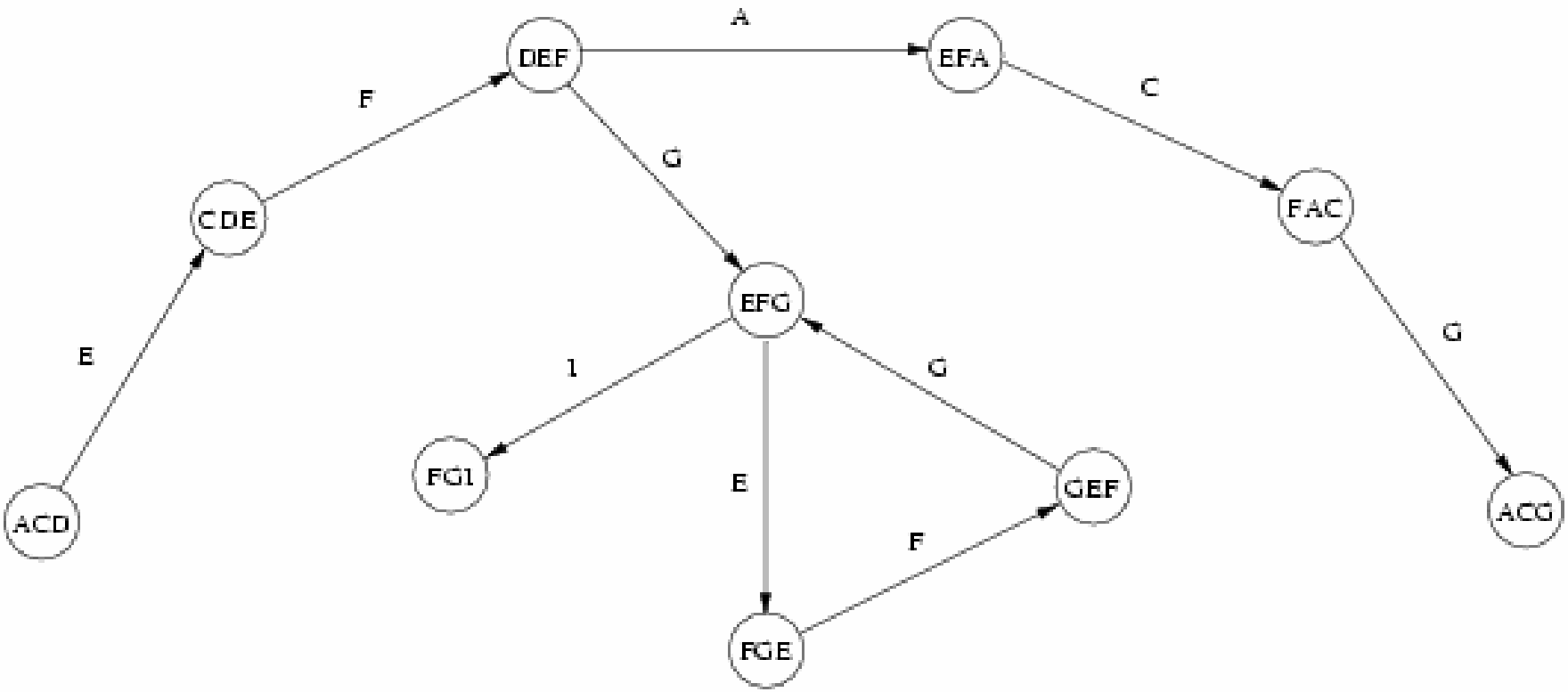
Eulerian tour constructs de Bruijn sequence.

# Sequence Database Compression

Construct sequence database / superstring that is

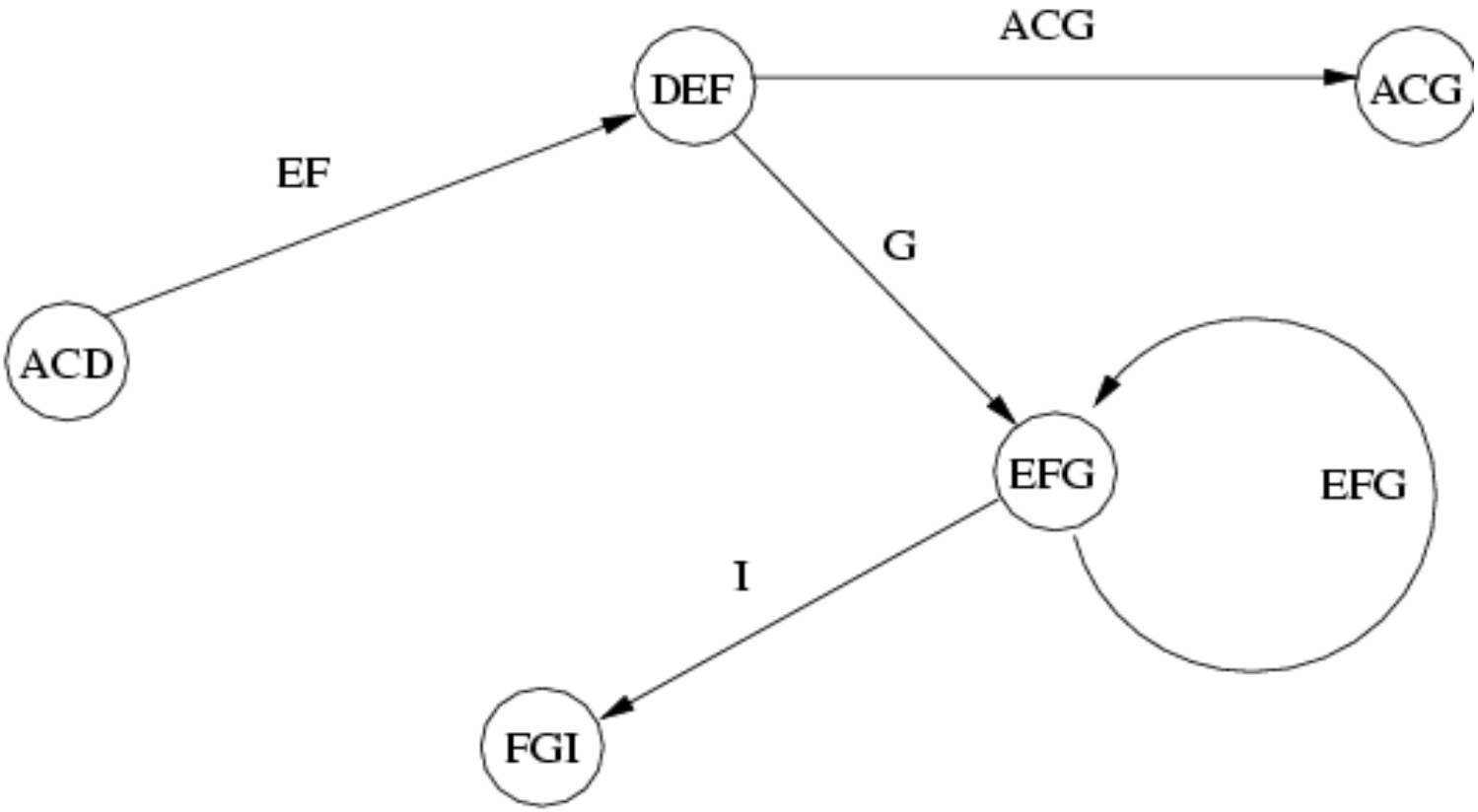
- Complete
  - All 30-mers are present
- Correct
  - No other 30-mers are present
- Compact
  - No 30-mer is present more than once

# SBH-graph




ACDEFGI, ACDEFACG, DEFGEFGI

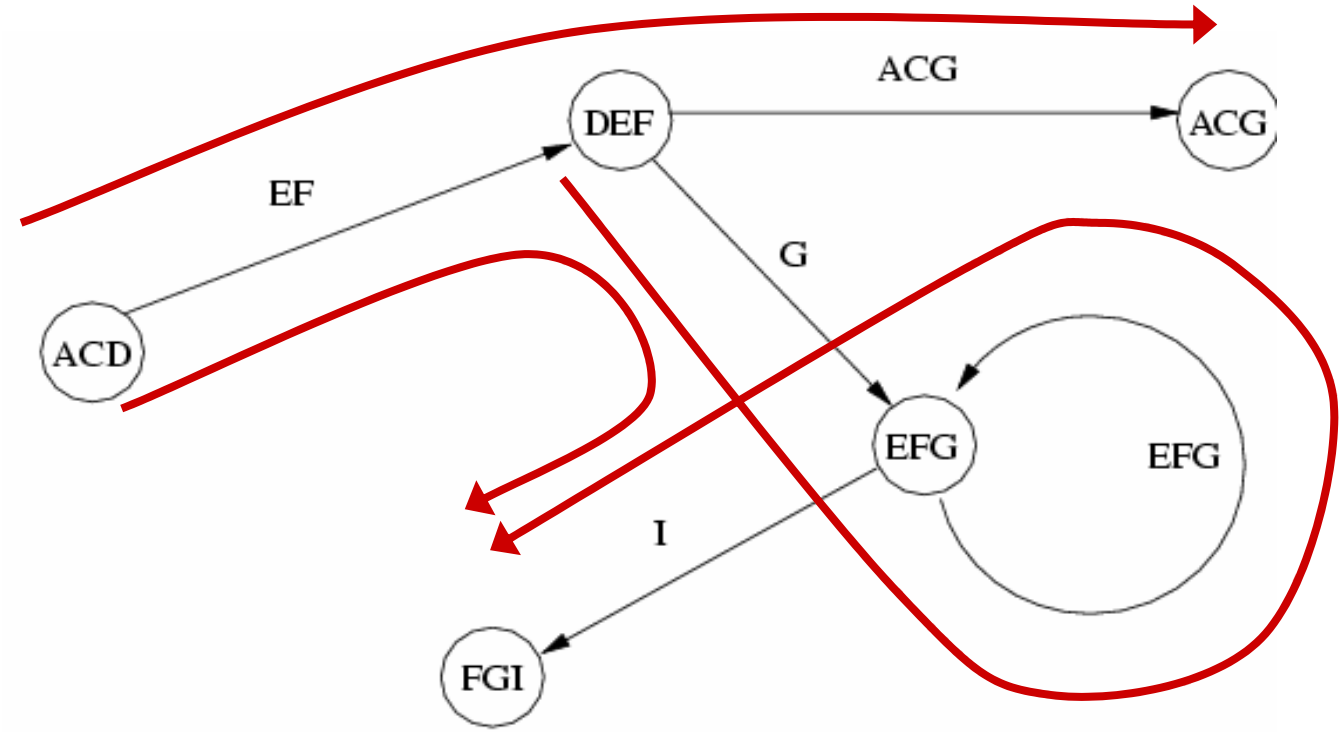
# Compressed SBH-graph




ACDEFGI, ACDEFACG, DEFGEFGI

# Sequence Databases & CSBH-graphs

- Original sequences correspond to paths




ACDEFGI, ACDEFACG, DEFGEFGI

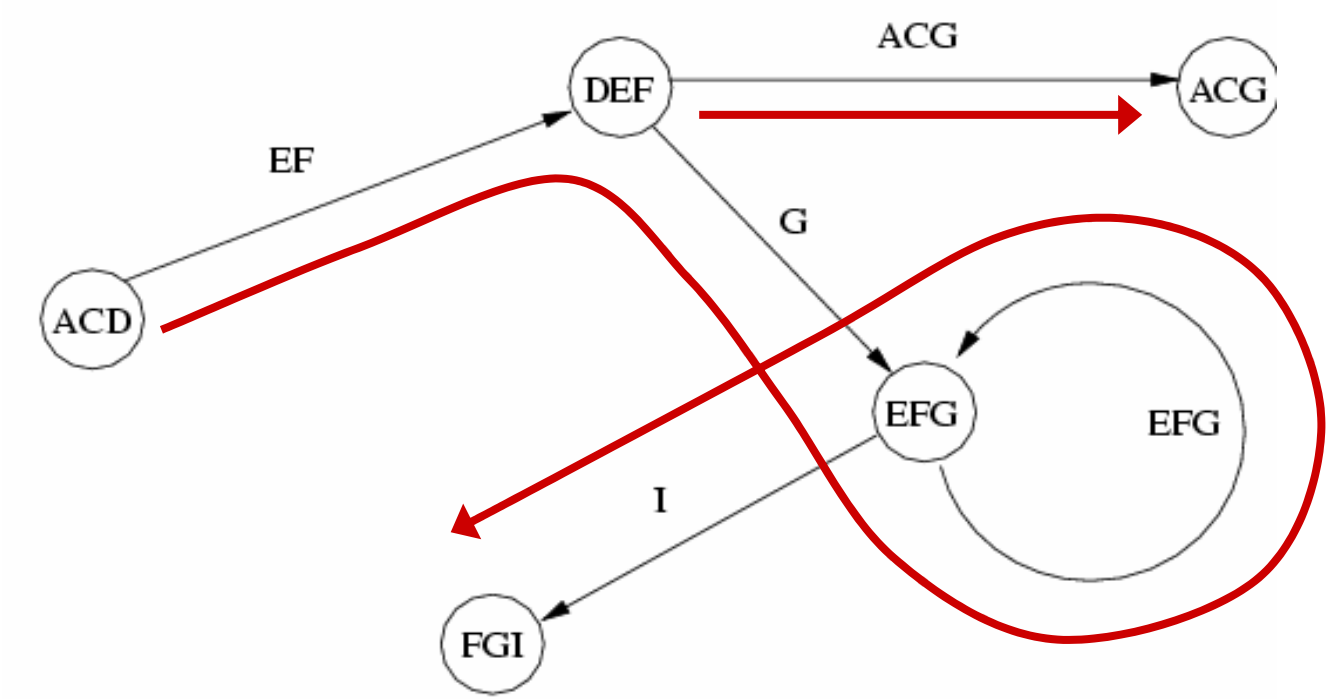
# Sequence Databases & CSBH-graphs

- Complete
  - All edges are on some path
- Correct
  - Output path sequence only
- Compact
  - No edge is used more than once

- $C^3$ Path Set uses all edges exactly once.

# Sequence Databases & CSBH-graphs

- Use each edge exactly once




ACDEFGEFGI, DEFACG

# Size of $C^3$ Path Set for k-mers

- Each path costs

  ($k$-1)-mer + path sequence + EOS

- Sequence database with $p$ paths

  $N_k + p\,k$

- Minimize sequence database size by minimizing number of paths
  - subject to $C^3$ constraints

# Best case senario…

…if CSBH-graph admits an Eulerian path.

Sequence database size

$$(k\text{-}1) + N_k + 1$$

How many paths are required if the CSBH-graph is not Eulerian?

# Non-Eulerian Components

- Net degree
  - b(v) = # in edges - # out edges
- Total degree surplus
  - $B_+ = \sum_{b(v)>0} b(v)$
- For each path
  - Start node's net degree +1
  - End node's net degree -1
  - Otherwise, net degree: no change
- To reduce all nodes to net degree 0, must have at least $B_+$ paths.

# Components w/ $B_+(C) == 0$

- Balanced component must have Eulerian tour, so require exactly one path.
- m balanced components

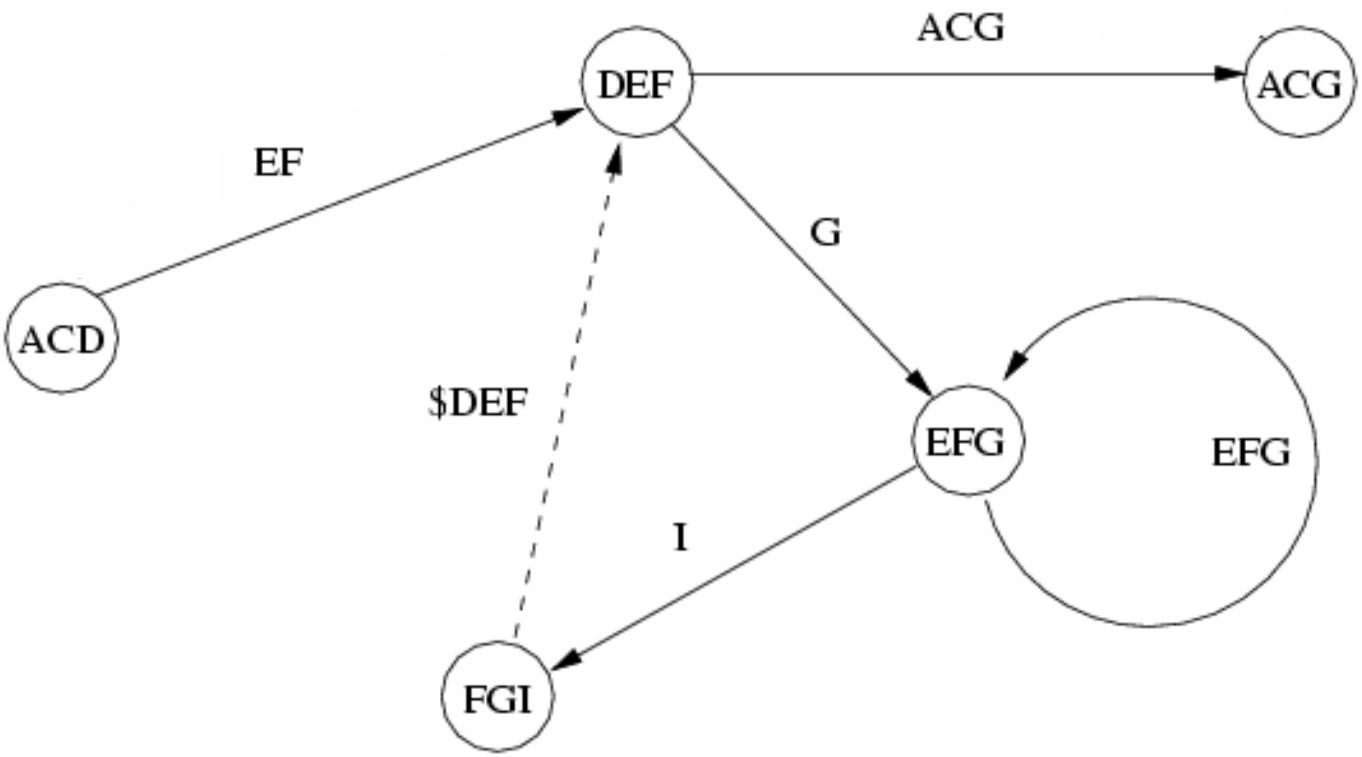
# # Paths Lower Bound

The $C^3$ path set must contain at least $B_+ + m$ paths.

This lower bound is achievable!

Just add $(B_+ - 1)$ “restart” edges to non-Eulerian components

# Achieving Path Lower Bound



# *k*-mer superstrings

- Solution is optimal, for $C^3$ constraints
- Polynomial time algorithm in length of original sequences

- General superstring problem
  - Requires completeness only
  - NP-hard [ Garey & Johnson ’79 ]
  - Approximable within a factor of 2.5
  - MAX-SNP hard

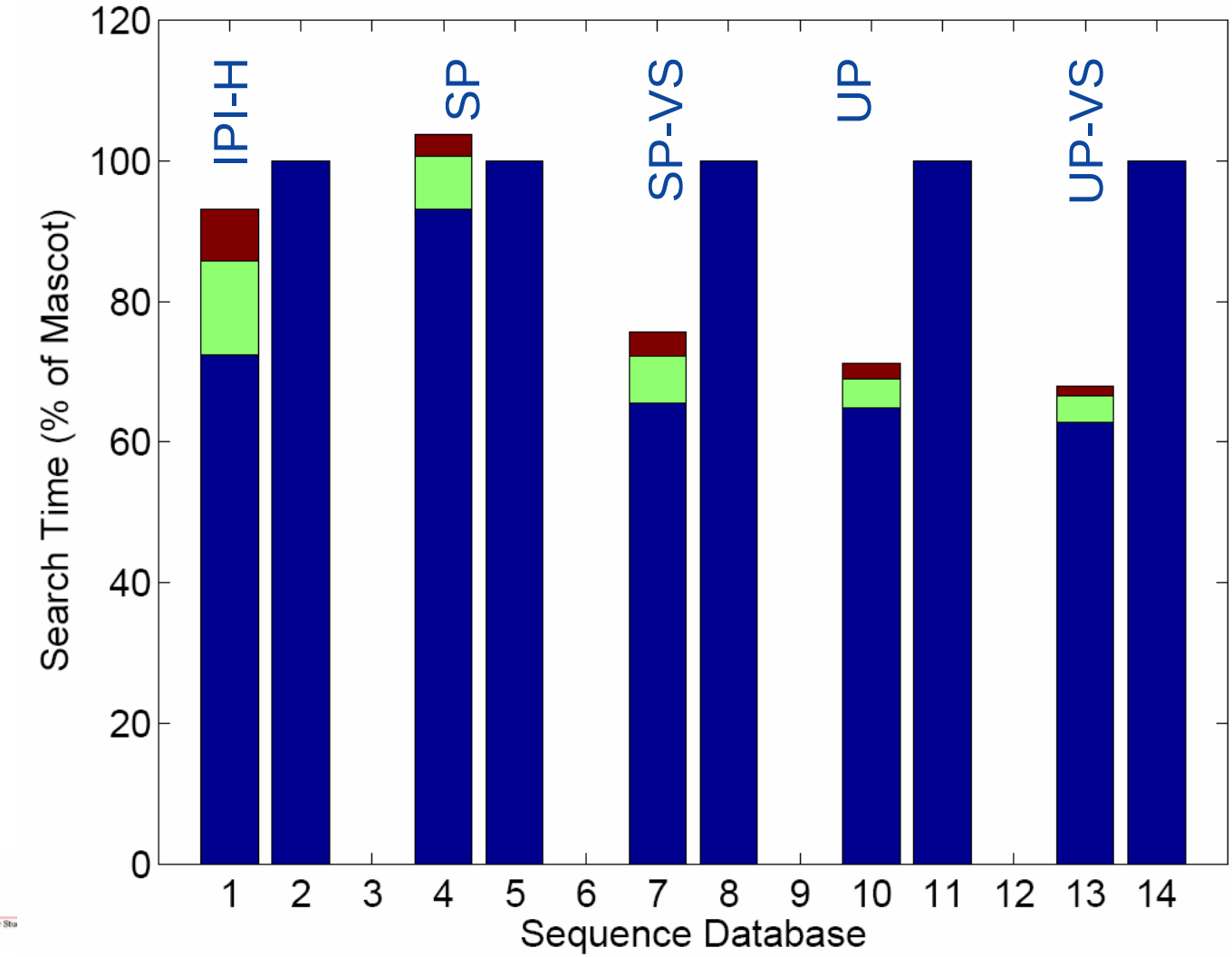
# AA Sequence Databases

| Sequence Database | Sequence Length | Distinct 30-mers | Overhead |
|---|---|---|---|
| IPI-HUMAN | 20358846 | 12115520 | 68% |
| IPI | 54145883 | 29769766 | 81% |
| Swiss-Prot | 56454588 | 44374286 | 27% |
| Swiss-Prot-VS | 89541275 | 45307827 | 97% |
| UniProt | 472581860 | 274510105 | 72% |
| UniProt-VS | 506796094 | 275391669 | 84% |
| MSDB | 481919777 | 276523755 | 74% |
| NRP | 495502241 | 283160529 | 75% |
| NCBI-nr | 619132252 | 378721915 | 63% |
| UnionNR | 674700840 | 385369671 | 75% |
| Union | 2157353500 | 385369671 | 460% |

# Minimum Size $C^3$ Sequence Database

| Sequence Database | $C^3$ 30-mer Enumeration | Overhead | Compression | Compression Bound |
|---|---|---|---|---|
| IPI-HUMAN | 13854679 | 14.35% | 68.05% | 59.51% |
| IPI | 37961385 | 27.52% | 70.11% | 54.98% |
| Swiss-Prot | 52662145 | 18.68% | 93.28% | 78.60% |
| Swiss-Prot-VS | 54534356 | 20.36% | 60.90% | 50.60% |
| UniProt | 337119564 | 22.81% | 71.34% | 58.09% |
| UniProt-VS | 338890778 | 23.06% | 66.87% | 54.34% |
| MSDB | 342924164 | 24.01% | 71.16% | 57.38% |
| NRP | 351600578 | 24.17% | 70.96% | 57.15% |
| NCBI-nr | 463517034 | 22.39% | 74.87% | 61.17% |
| UnionNR | 473665310 | 22.91% | 70.20% | 57.12% |
| Union | 473665310 | 22.91% | 21.96% | 17.86% |

# Relative Search Time



# Constraint Relaxation

- Why insist on compactness?
  - What about 29-mers?
- Can we compress still further?
  - Complete, Correct ($C^2$)
  - Use edges more than once, if helpful!
- How could this possibly help?

# $C^2$ Superstring

- Sequence set with *p* paths

$$N_k + p\,k$$

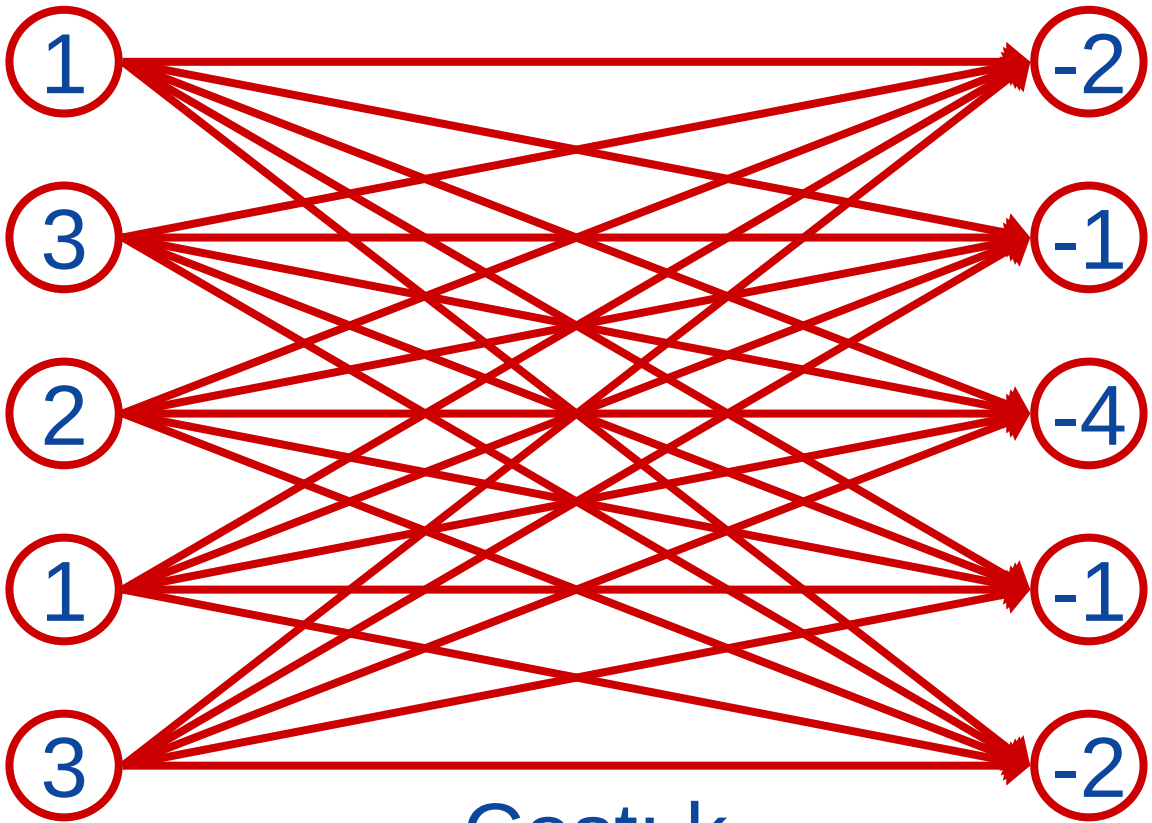
- Costs *k* to use restart edges
  - Restart edges from nodes v s.t. b(v) > 0 to v s.t. b(v) < 0
  - Reuse edges instead!

    …provided the path length is < k
- Transportation problem!

# $C^2$ Superstring

S = {v:b(v)>0}

T={v:b(v)<0}

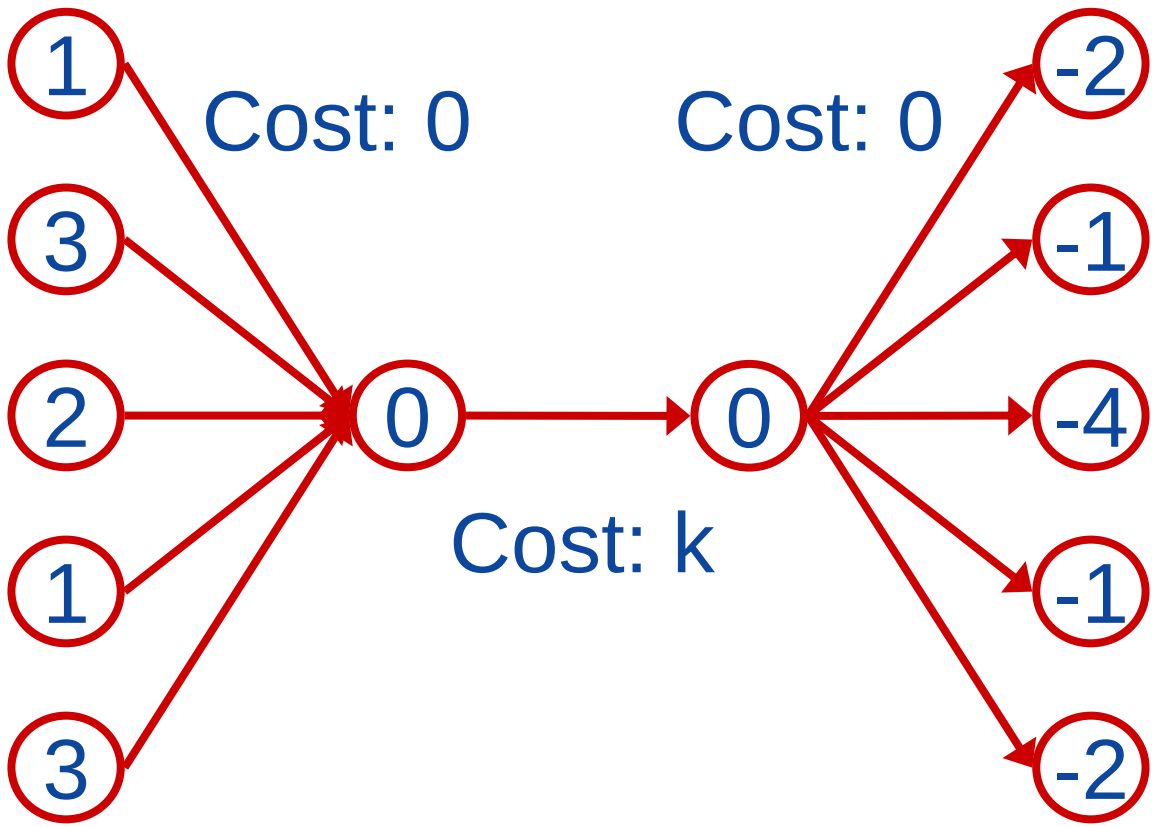



Cost: k

# $C^2$ Superstring

S = {v:b(v)>0}

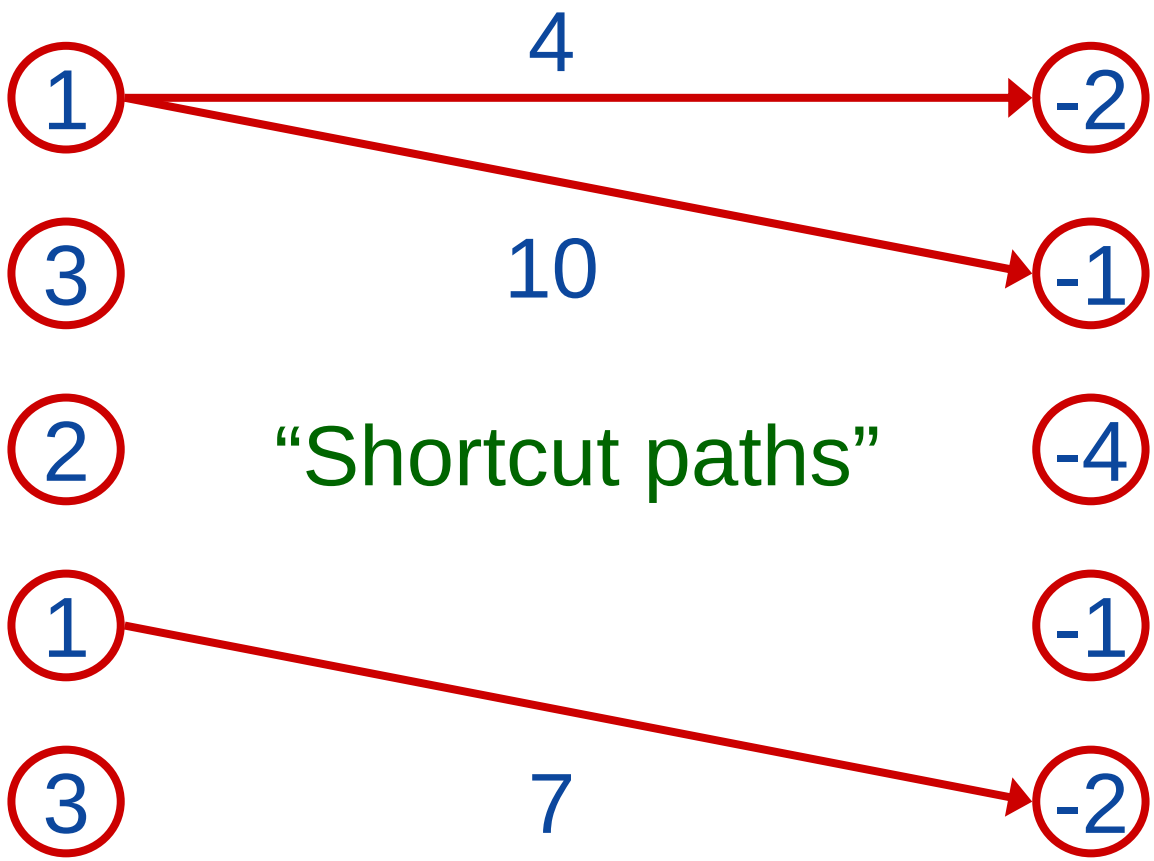
T={v:b(v)<0}



# $C^2$ Superstring

S = {v:b(v)>0}

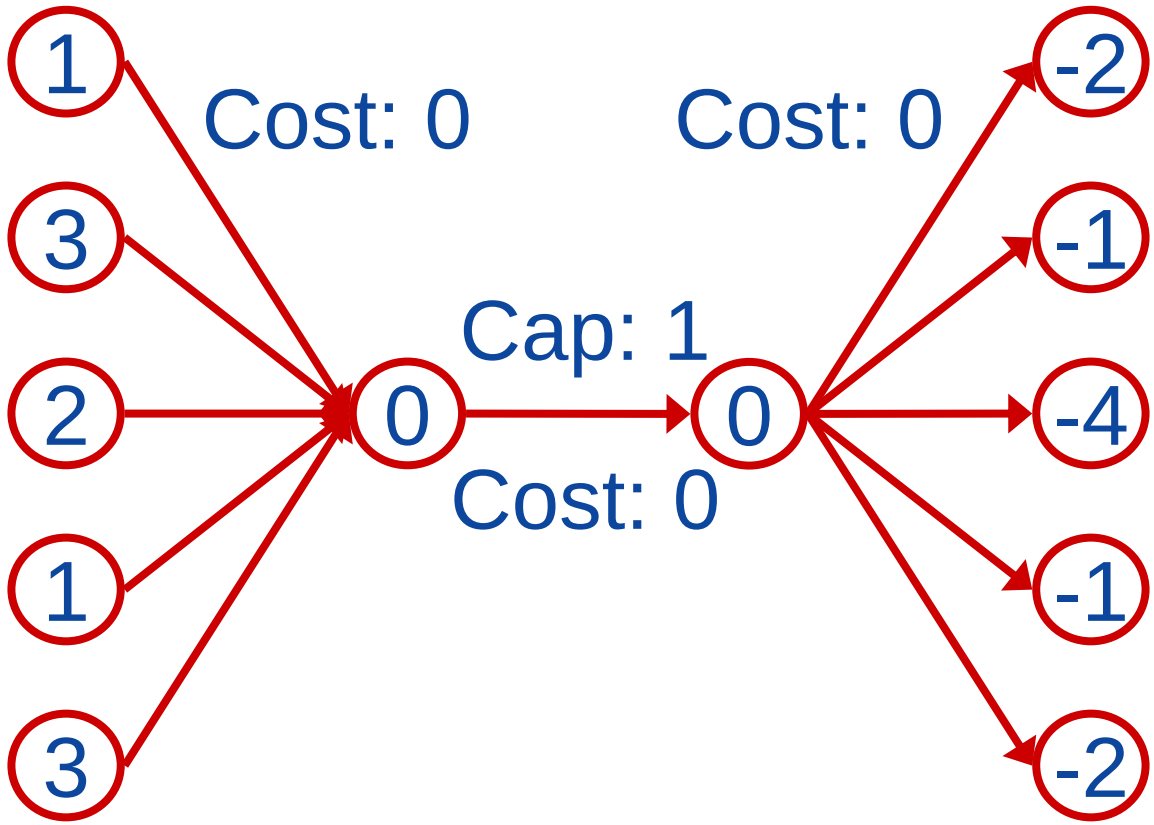
T={v:b(v)<0}



# $C^2$ Superstring

S = {v:b(v)>0}

T={v:b(v)<0}



# $C^2$ Superstring

| Sequence Database | $C^3$ Superstring Overhead | $C^3$ Superstring Compression | $C^2$ Superstring Overhead | $C^2$ Superstring Compression | Compression Bound |
|---|---|---|---|---|---|
| IPI-HUMAN | 14.98% | 67.36% | 13.25% | 66.35% | 58.58% |
| IPI | 27.65% | 68.84% | 20.67% | 65.08% | 53.93% |
| Swiss-Prot | 18.78% | 91.87% | 14.62% | 88.66% | 77.34% |
| Swiss-Prot-VS | 20.50% | 59.87% | 15.98% | 57.62% | 49.68% |
| MSDB | 23.78% | 71.73% | 18.40% | 68.61% | 57.95% |
| UniProt | 22.76% | 73.90% | 17.43% | 70.76% | 60.20% |
| UniProt-VS | 22.98% | 69.59% | 16.65% | 64.25% | 56.59% |
| NRP | 24.17% | 71.61% | 18.64% | 68.42% | 57.67% |
| NCBI-nr | 22.69% | 74.85% | 17.56% | 71.71% | 61.01% |

# Extensions and Futher Work

- Better compression
  - Enumerate tryptic peptides only
  - Relax correctness constraint
- Other uses of CSBH graphs
  - Compact representation of mer counts
  - Implicit set operations on mers
  - Structural graph properties

# Thanks

- Informatics Research @ ABI & Celera
  - Ross Lippert, Clark Mobarry, Bjarni Halldorsson
- UMIACS @ University of Maryland, CP
  - V.S. Subrahmanian, Fritz McCall, Doan Pham