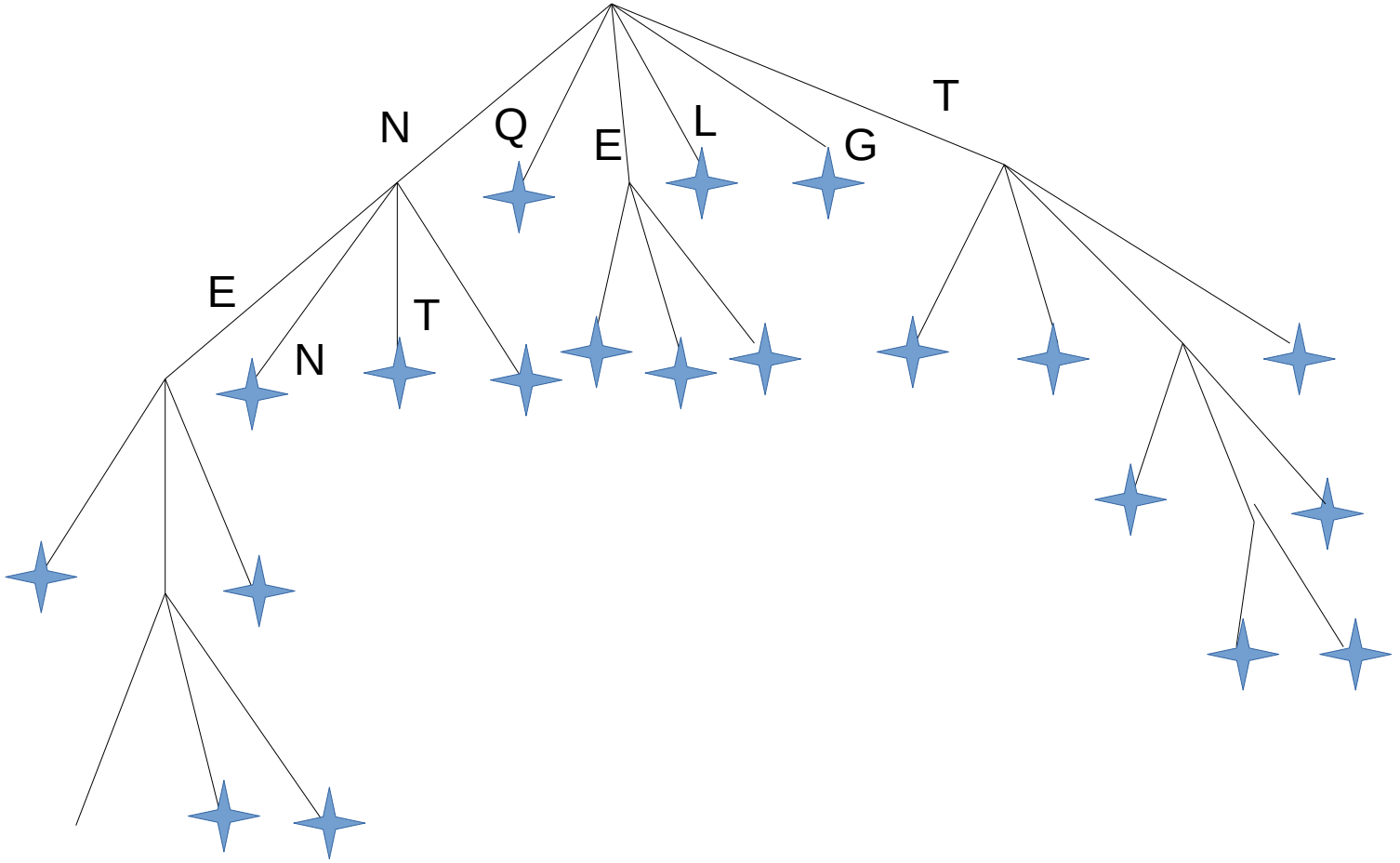
CMSC423

Chapter 4 – Proteomics/mass-spectrometry
Solving imperfect spectra

Summary

- Create table of peptides of increasing length
- Check each peptide's LINEAR spectrum against experimental spectrum (check for containment)
- Discard peptides with masses not in experimental spectrum
- Stop when one peptide has CIRCULAR spectrum matching experimental spectrum

Branch and bound



Stop and think!

Without the "bound" step, how big is the search space for a peptide of length k?

Dealing with errors

- Even one error in experimental spectrum can "disqualify" correct answer
- Remind you of anything you've seen?

- Instead of "match/no match" look for score of match: # of masses in theoretical spectrum found in experimental spectrum

- Why not also account for # of masses in experimental spectrum not found in theoretical spectrum?

Matching spectra (with errors)

Peptide: GASP 57-71-87-97

Theoretical:	G	A	S	P	GA	PG	AS	SP	GAS	PGA	SPG	ASP	**GASP**
	57	71	87	97	128	154	158	184	215	225	241	255	**312**

Theoretical: (with errors)	G	A	*S*	P	GA	PG	*AS*	SP	GAS	PGA	SPG	ASP	**GASP**
	57	71	*89*	97	128	154		184	215	225	241	255	**312**

Partial peptides (bold if matching):

GAS:	G	A	S	GA	AS	GAS
	57	**71**	87	**128**	158	**215**

Score = 4 (matching masses)

APS:	A	S	P	AP	PS	APS
	71	87	**97**	168	**184**	**255**

Score = 4 (matching masses)

New algorithm

- Don't assume experimental spectrum is perfect
- Generate all peptides of length 1
- Keep the best matching one
- Extend it by one amino acid
- Keep the best matching one
- Repeat...

Any issues?

New algorithm

- Don't assume experimental spectrum is perfect
- Generate all peptides of length 1
- Keep the best matching ones (top of the leaderboard)
- Extend it by one amino acid
- Keep the best matching ones
- Repeat...

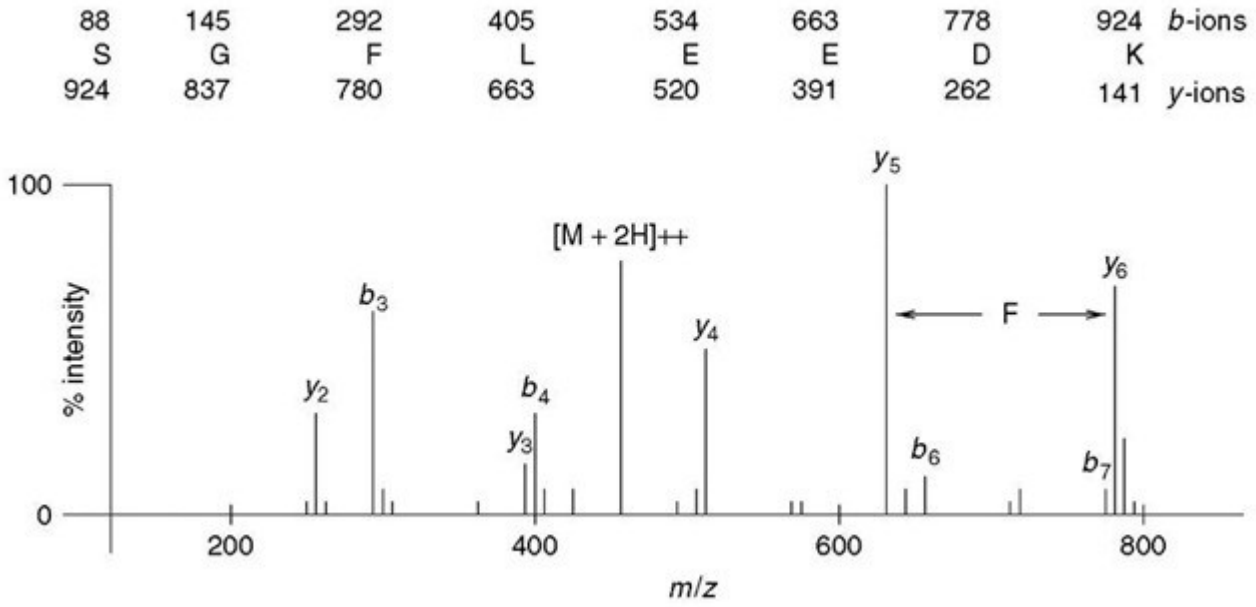
Stop and think!

- How does the size of leaderboard (# of top matches) impact:
  - run time?
  - ability to find the correct peptide?

- Also: when do you stop (if finding match not guaranteed)?

What if you don't know weights?

- Easy – infer from experimental spectrum



ELVISLIVES

E = ELVISLIVES – LVISLIVES
E = SELVISLIVE – SELVISLIV
E = ELV – LV
E = LIVE – LIV
...

The most frequent small differences are the amino acid masses

(spectral convolution)

Full algorithm

- Infer amino-acid masses from spectrum (if you cannot trust your database)
- Run leaderboard algorithm using the inferred mass "dictionary"