ILSI Health and Environmental Sciences Institute

Modeling local alignments within allergens and celiac proteins

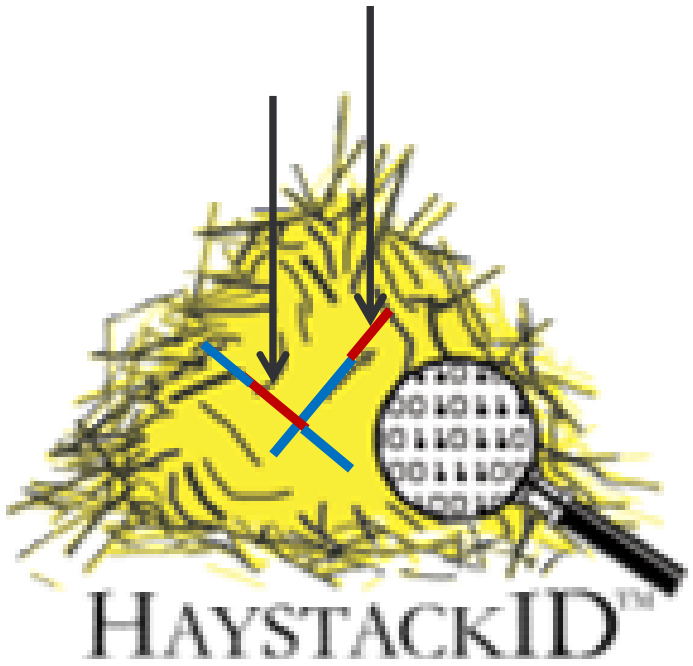
Rome, Italy October, 2016

Scott McClain, Ph.D.
Syngenta Crop Protection, LLC
Research Triangle Park, NC, USA

Bioinformatic sequence matches – BLAST and FASTA

- Designed to find local alignments across two or more sequences.

This concept is like not only finding the needle in a haystack, but finding several needles all at once, and then finding the similarity that may be shared within each needle.

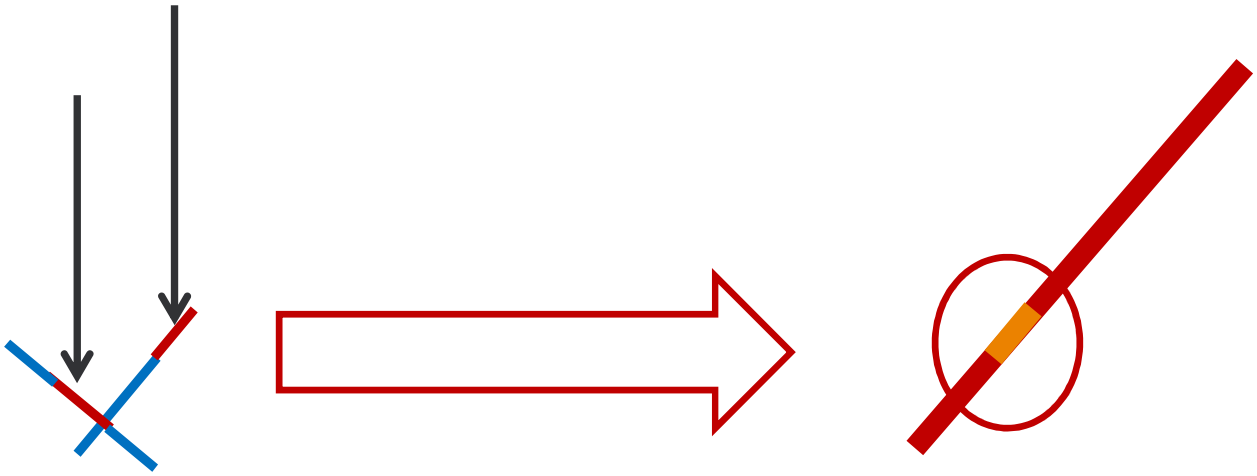


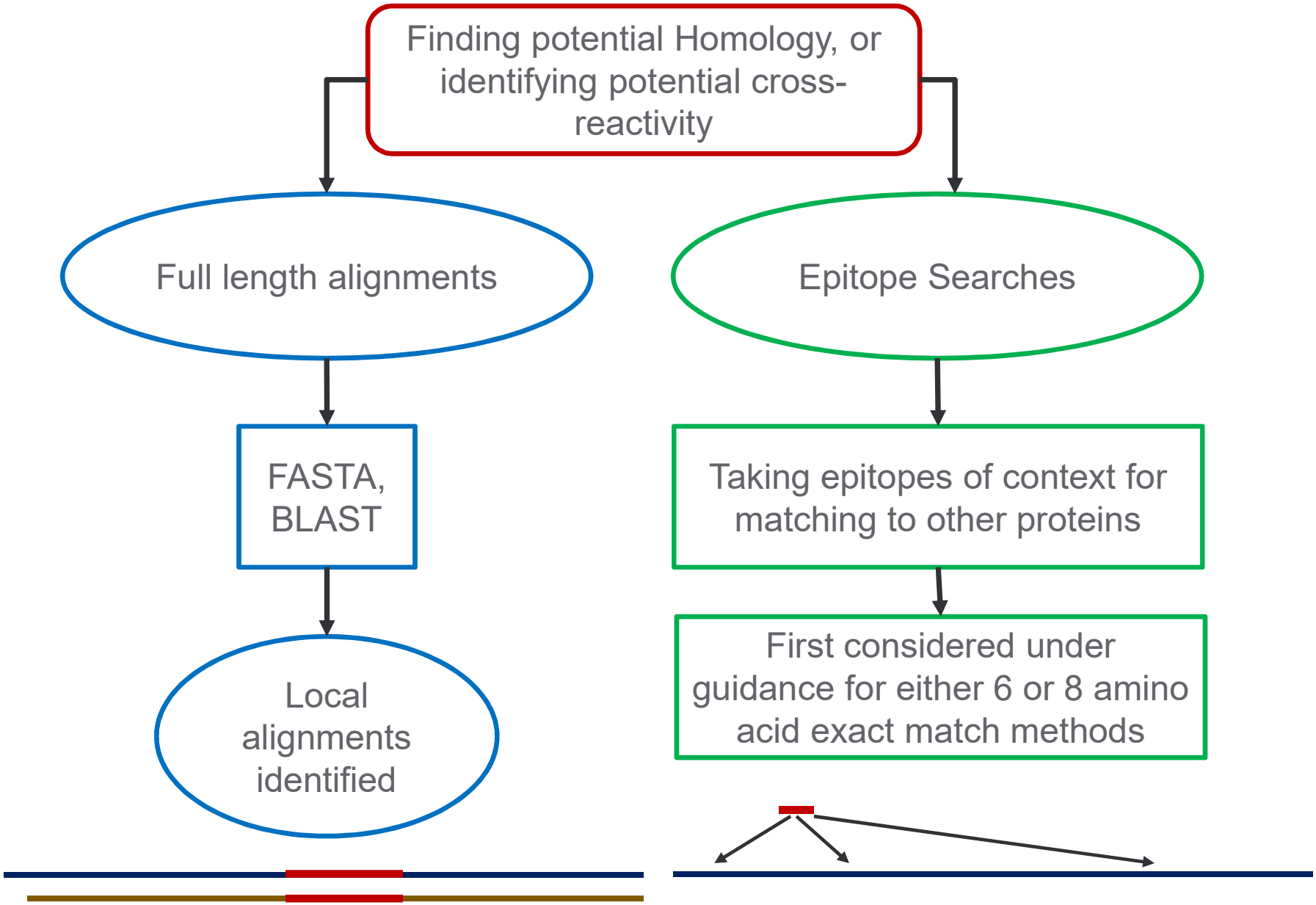
% Identity – this helps find the needles, but does little more. It does not help identify similarity based on interpretation of the secondary and tertiary structure.

Bioinformatics, Allergens and Celiac Peptides

- Question - How to test whether local-alignment informatics can find important allergen alignments?

- **Peptide matches are a Hyper-local view of similarity between proteins; i.e., examining a small portion of a larger, intact protein**
    - *The challenge; you expect limitations when the length of a peptides doesn't include representation of all 20 amino acids!*




Finding potential Homology, or identifying potential cross-reactivity
Full length alignments
Epitope Searches
FASTA, BLAST
Taking epitopes of context for matching to other proteins
Local alignments identified
First considered under guidance for either 6 or 8 amino acid exact match methods

Hypothesis – it is possible to better understand allergens and celiac proteins using informatics

- Basis for this:
  - Both, allergens and celiac proteins are special proteins
  - Allergens cross-react amongst themselves if closely related enough
  - Cross-reactivity relies on localized, immune-reactive sequence called epitopes
  - Epitopes function as either linear or discontinuous (or both?)
    - Note: discontinuous epitopes require a level of estimated, shared sequence that accommodates higher levels of protein organization, rather than just linear
  - FASTA can perform sophisticated non-linear estimates of shared identity and similarity that is appropriate for **allergens**

Epitopes are Specific, but *how* are they specific?

M R T G A D P Q N

Even at 9 amino acids – still not close to representing the 20 amino acids library

In first considering “small amino acid, exact matches”, the premise was that a match between a novel protein and an allergen could be a putative epitope

The putative hazard being that if it was an epitope match then the cross-reactive (health hazard) existed for allergen sensitive patients

First problem, for most allergens the epitopes are unknown

Second problem, allergens are “unique” in having IgE binding epitopes, but very short and **random** (8 or 9) amino acids sequences are NOT unique

What we now from studying 8-mer “epitopes” when treating them as a screen for allergen matches…

- From Hileman et al 2002 –
  - As exact matches between proteins is shortened from 8 to 6 amino acids, the utility in identifying allergens is degraded to the point of uselessness.

- From Silvanovich et al (2006)
  - At 6 aa, any matches between any protein and allergens would indicate that 60 – 80% of all proteins would share an exact match
    - Matches are random and do not indicate relevant similarity or indication of cross-reactivity
  - **Only at 8-9 aa** was there the ability to describe a unique match between a protein and a known allergen **without artefactual alignments based on database size, protein length and peptide length**.

Epitopes

- The conclusion is that short peptides when treated as single “query” sequences have virtually no statistical power to identify shared immune specific sequence between two or more sequences…
- If “epitopes” are biologically relevant when present in allergen sequences, AND allergens occur uniquely in restricted taxonomic groups…then what?
- Then, epitopes matter when
  - 1) they are known and they are identified as part of a whole protein known as an allergen…and, two or more are biologically relevant.
  - 2) the epitopes of an allergen occur uniquely as part of the larger contextual sequence that makes up the full-size protein
    - **This restricts identity of an epitope to both the protein and the organism in which it resides.**

Approach…

- …to use **biologically relevant epitope** informatics to drive towards a meaningful informatics threshold for similarity

A hypothetical sequence "spiked" with Bet v 1 discontinuous epitope residues; total length = total 160 aa.

Underlined letters are the Bet v 1 epitope residues.

PVEVKNYLSSIIIVLDISTSSFTFTIIKIPSRIWDCWKVPT**ENIEGNGGPGTIKKISFPEGLPFKYVK**

DRVDEVDHTNFKYNYSVIEGGPIGDTLEKSRVEVTSTGTACSCTHSITTHMINYSIHINTNLENNSLT

PTLLHLVTMFLNSSSFPDLVHNLH

Note, I'm using a "region" and not just the epitope residues…

What to pay attention to…

- 1) At which threshold does E-value distinguish known cross-reactive homologues in other species?
  - Can satisfy this empirically, because with well studied allergens, there is pretty good clinical data to understand cross-reactive homologues and determine threshold using E-value.

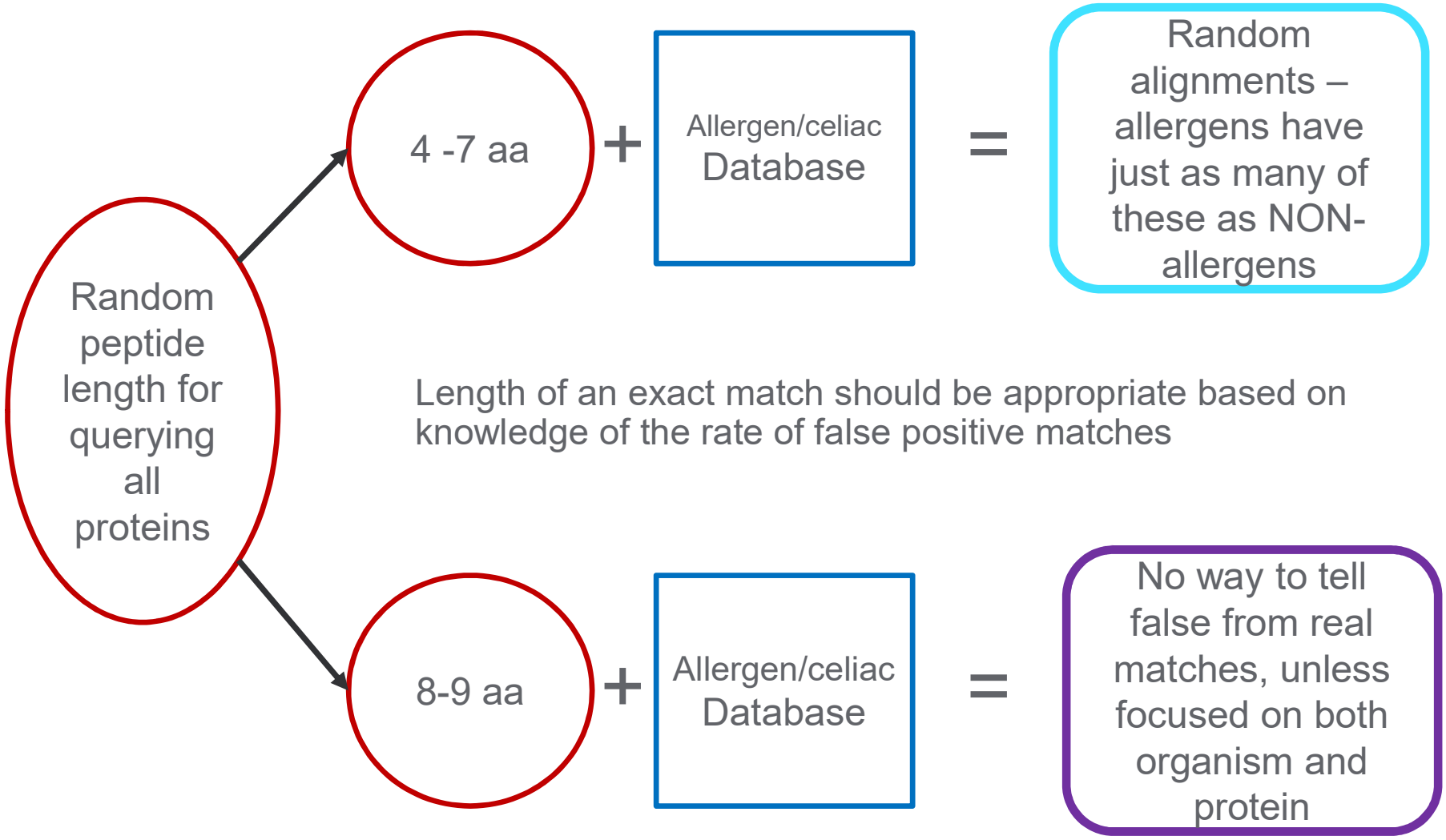
Bet v 1 – the whole 56 AA region

Order of decreasing significance

Allergen		Species	% Identity	Alignment Length	E-value
PRP-Like Protein	302379157	*Daucus carota*	34.6	78	6.60E-06
PRP-Like Protein	302379147	*Daucus carota*	33.3	78	1.10E-05
PRP-Like Protein	302379149	*Daucus carota*	33.3	78	1.10E-05
PRP-Like Protein	302379155	*Daucus carota*	32.1	78	2.80E-05
Pathogenesis-Related Protein-Like Protein 1	19912791	*Daucus carota*	33.3	78	2.80E-05
Major Allergen Api G	14423646	*Apium graveolens*	32.9	76	2.90E-05
PRP-Like Protein	302379151	*Daucus carota*	32.1	78	5.30E-05
PRP-Like Protein	302379153	*Daucus carota*	32.1	78	5.30E-05
Cytokinin-Specific Binding Protein	4190976	*Vigna radiata*	31.3	96	3.00E-03
Per A 4 Allergen	60678787	*Periplaneta americana*	39.6	48	6.30E+00
Vacuolar Serine Protease	12005497	*Penicillium oxalicum*	35.8	53	7.50E+00

Fairly clear line in the sand with allergen – ***Question***, does this line the same for other large groups?

Where is the tie-in to Celiac peptides?

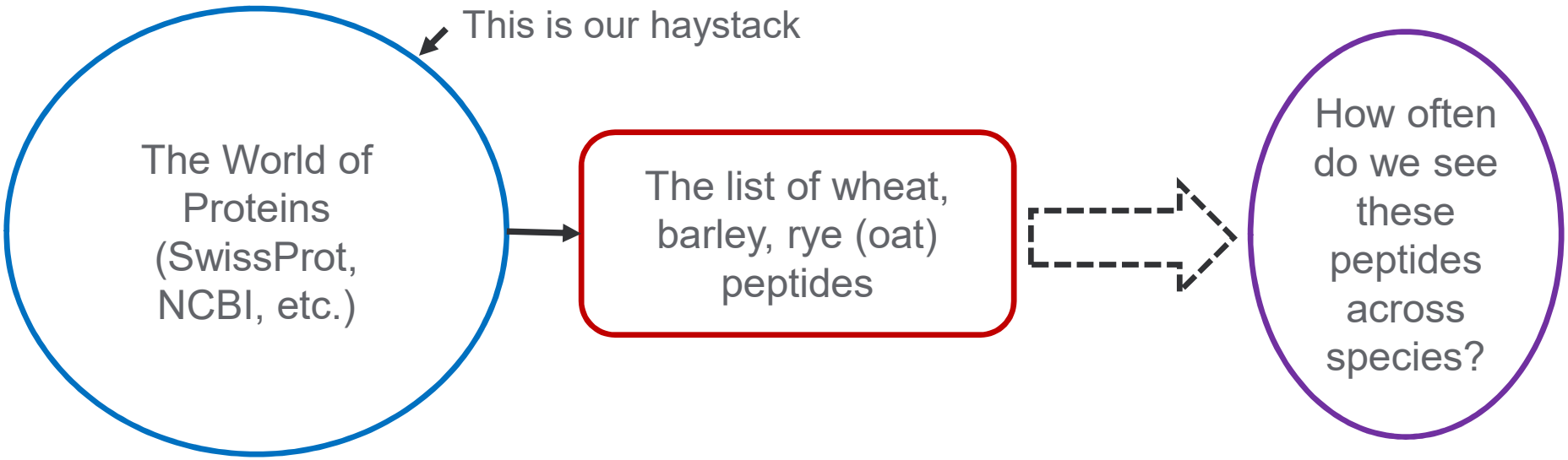


How can we show informatically that celiac peptides could offer a way to screen all proteins for celiac risk?

As has been shown for “8-mer” matches under the premise of “finding” epitopes, we can screen for how often the known celiac peptides occur in organism/proteins not known to cause celiac…

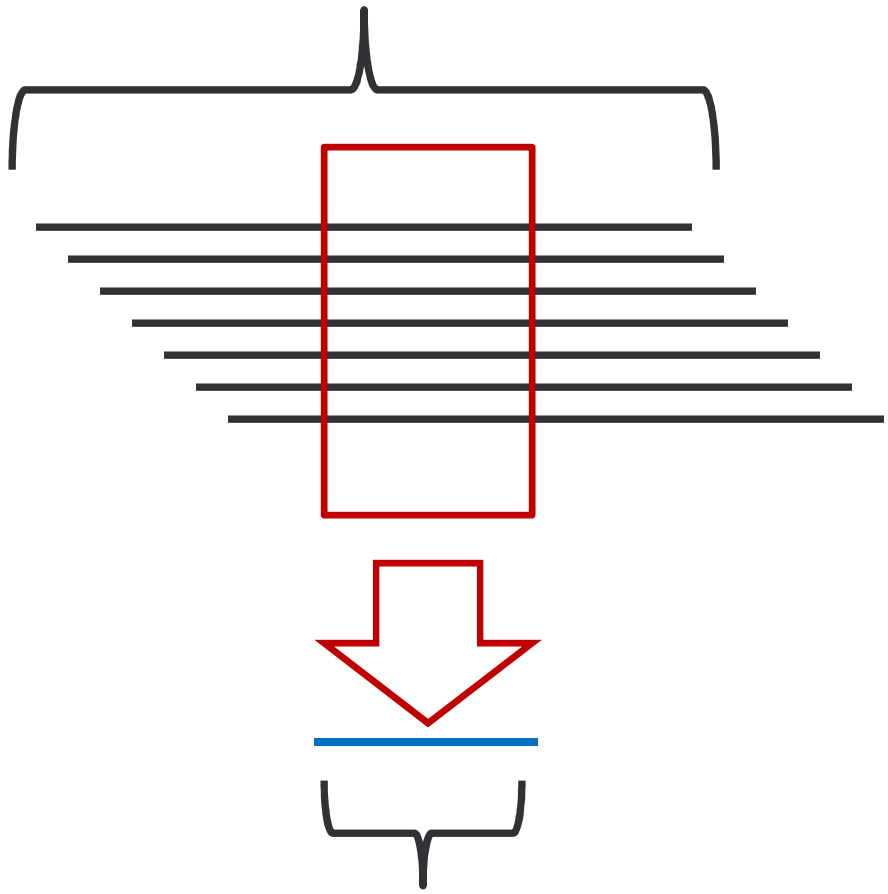
The reason to test this premise is that to date, only wheat, barely and rye (oat) have substantial clinical evidence for celiac disease.

The underlying premise - there is uniqueness to the organism and the specific sequence of glutenins and gliadins that is not found in other organisms and their proteins.



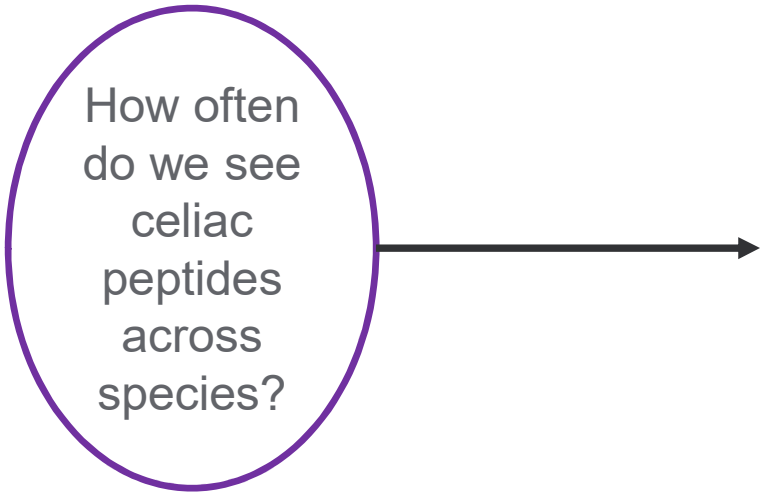
We know which 9 amino acid peptides exist as T-cell epitopes

9 amino acid common length



4 amino acid common length if variation is allowed

Testing the theory of exact matches before employing it as a safety screen…



If these peptides are observed in species and proteins *not associated with celiac disease* then this is our "false negative" rate.

In other words, the rate at which we observe celiac peptides outside the species range of wheat, barely and rye (oat) will indicate the statistical utility of screening non – wheat, barley and rye (oat) proteins for any matches…

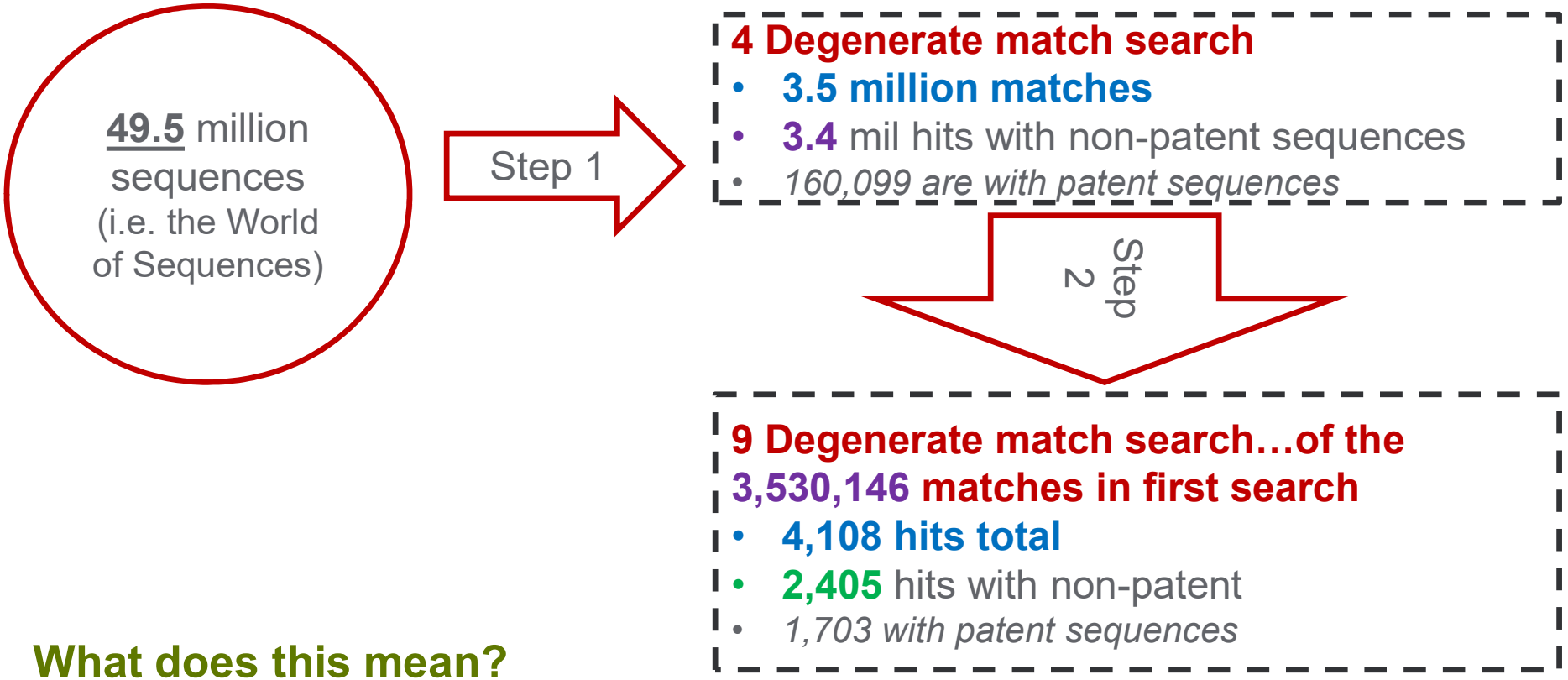
Degenerate "pattern matching" for the 4 peptide motif

Q/E-X1-P-X2_motif	**[QE] [LQFS] P [YFAVQ]**

This is what we'll look for first

First, look at all the available proteins…

Take the degenerate list of peptides with the variances observed in those peptides…i.e., allowing for mis-matches to account for the variation



What does this mean?

False positive rate = (1 - (**2405**/3,370,047))*100%)

99.93% of degenerate 4 amino acid peptide hits are with proteins that likely have no association with CD

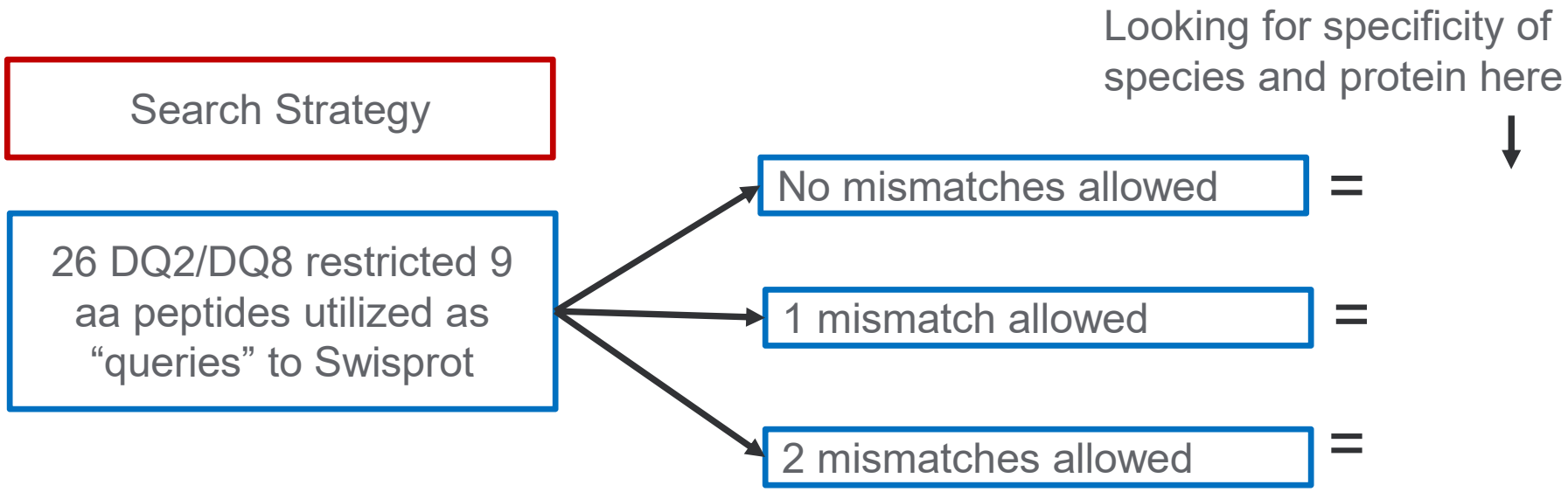
Data courtesy of Andre Silvanovich

Checking to see how often other species are identified by the celiac peptide list…

Data from SwissProt.

Summary of the **4 amino acid degenerate search**…similar outcomes as with previous data…widely dispersed identification of matches with thousands of proteins from **species other than wheat, barley and rye.**

Performing a 9 amino acid celiac peptide match/search…



Data courtesy of Elda Posada Campos

Results from 9 aa search of SwissProt

Mismatch = 0. Observe 90 matches. Found in proteins belonging to the expected species: Triticum aestivum (wheat), Hordeum vulgare (barley) and Avena sativa (oat).

Possible “mass action” impact from multiple proteins that trigger celiac disease, not just one protein

Mismatch = 1

Species	Common name	Number of alignments
Arabidopsis thaliana	Mouse-ear cress	4
Avena sativa	**Oat**	**6**
Candida albicans	Yeast	4
Dictyostelium discoideum	Slime mold	32
Drosophila melanogaster	Fruit fly	4
Hordeum vulgare	**Barley**	**77**
Homo sapiens	Human	15
Kluyveromyces lactis	Yeast	1
Magnaporthe oryzae	Rice blast fungus	2
Rattus norvegicus	Rat	4
Triticum aestivum	**Wheat**	**271**

66 others

Data courtesy of Elda Posada Campos

Results continued…

- **Mismatch = 2**. We observe 2,824 matches. 43 times more matches other than wheat. barley and rye (oat). **No useable specificity identified.**

- Focusing on one, non celiac-associated species…
  - Potato = 39 alignments. 18 of those 39 carry the Q/E 4 aa motif
  - If we back up and examine potato thoroughly using the whole genome with the 4 aa motif, but not allowing mismatches, there are **4,769** gene products (proteins) with alignments.

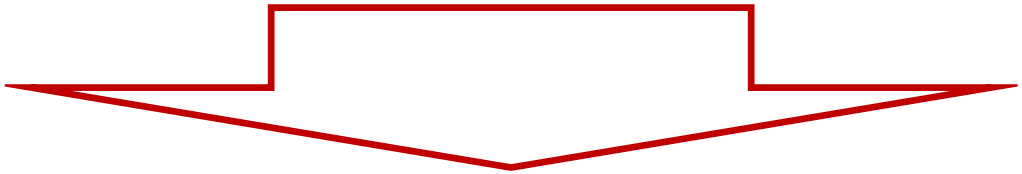
- The ability to distinguish celiac proteins with mismatches allowed is not possible.
- The ability to distinguish celiac peptides using a degenerate 4 aa peptide match also fails.

Conclusions…

- As in the past, random single small peptides are not long enough to act as a stand-alone distinguishing feature of allergens.
  - Known IgE binding epitopes on known allergens are flags for specific allergens in some cases, but they are not a screening tool without this knowledge prior to searches of allergens.
  - In other words, **random short peptides** are not a feature of allergens that can be used to screen for similarity between two proteins and assuming that they represent epitopes.
- Many epitopes together can be enough sequence information to model for an informatics threshold to identify homologous proteins that are allergens, but varies with length of the allergens and the number of epitopes.
  - Can be used to model thresholds for similarity metrics (E-value)

Are Celiac-associated peptides unique – do they behave in informatics like allergen epitopes?

- The better way to ask this question is…How could celiac peptides be unique?
- The short answer is the same as for IgE-binding epitopes…
  - Celiac peptides are unique in association with only those proteins and those organisms known to possess clinical risk of gluten-associated enteropathy.



Therefore, the combination of the

1) peptides themselves,
2) contextual protein-specific sequence in which the peptides reside,
3) specific protein structure/function unique to the organism, and
4) specific exposure context (dose and thresholds) is what describes celiac risk (hazard plus exposure).

Conclusions, continued…

- Celiac summary conclusions
  - The evidence shows that indeed, only those celiac species of interest contain the 9 amino peptides.
  - Peptides work as screening tools only under the following conditions
    - When they are 9 or longer and specific for celiac species
    - When the match is exact
  - A safety screen would only make sense when applied to protein specific peptides of appropriate length (i.e., 9 aa exact matches for known celiac peptides).
  - There is no indication that the organism itself is of risk other than as a whole food.
  - When examining individual proteins for celiac similarity, it is the protein, not the organism that carries the risk of sharing similarity with another protein.

Discussion/Q & A

- Thanks!