Database of Celiac Peptides; What it is and how it is Best Used

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

- Are celiac peptides unique and do they comprehensively tag/flag the offending proteins?
- Corollary is similar to 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 [*PEER REVIEW: EFSA, COMPARE (2019)*].

Therefore, the combination of the following define celiac risk (hazard ***MULTIPLIED*** by exposure).

1) HAZARD – known CD-associated peptides,
2) HAZARD - contextual protein-specific sequence in which the peptides reside (i.e. the rest of the gluten protein; there is a whole molecule impact on digestion and tTG specific activity,
3) HAZARD - specific protein structure/function unique to the CD organism, and
4) EXPOSURE - specific exposure context (dose and thresholds) is what describes celiac risk (hazard plus exposure).

The database teaches how to find those rare or special sequences in the World of Proteins

Putative hazard hypothesis: supports using a celiac peptide database to identify putative CD peptides in non-celiac proteins:

- Non-celiac proteins may contain the "active" or peptide portion of a celiac protein.

Putative Methodology for Safety: for proteins that may enter the food chain in a new or different way – I.E., GM protein traits (or any protein of interest [**POI**]):

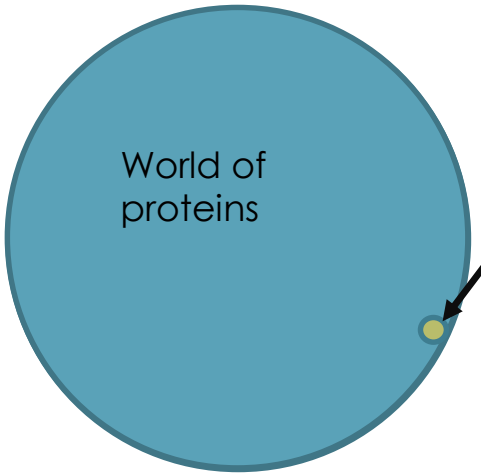
- Accurate, specific and sensitive sequence level screening method

Validation: sequence comparative methods rely on two components ***that must be paired***:

1. A qualified database of Celiac Peptides
2. Validated methods that can identify with high sensitivity and specificity – *i.e., all the sequences in the database* ***can be identified***.

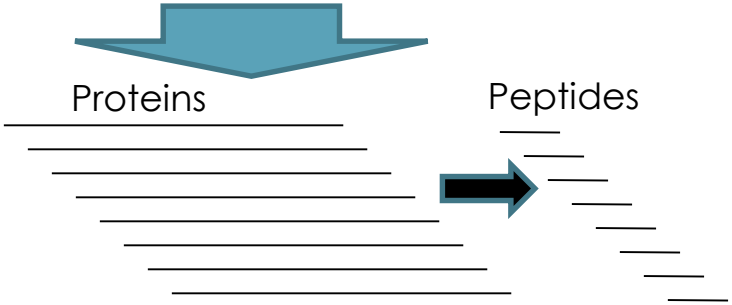
Building the right bioinformatics with the right Database

Hence – **Tier 1** begins with **source organism** in terms of understanding Risk



Celiac proteins

Clinical Science defines these by Source organisms that produce clinical symptoms in those patients who consume food – i.e., ***the agent is not a pure gluten, but a mix of proteins***



Availability of a list of peptides derived from large proteins that make up what we call a **Celiac Peptide Database**

TIER II

How is hazard characterized with this database?

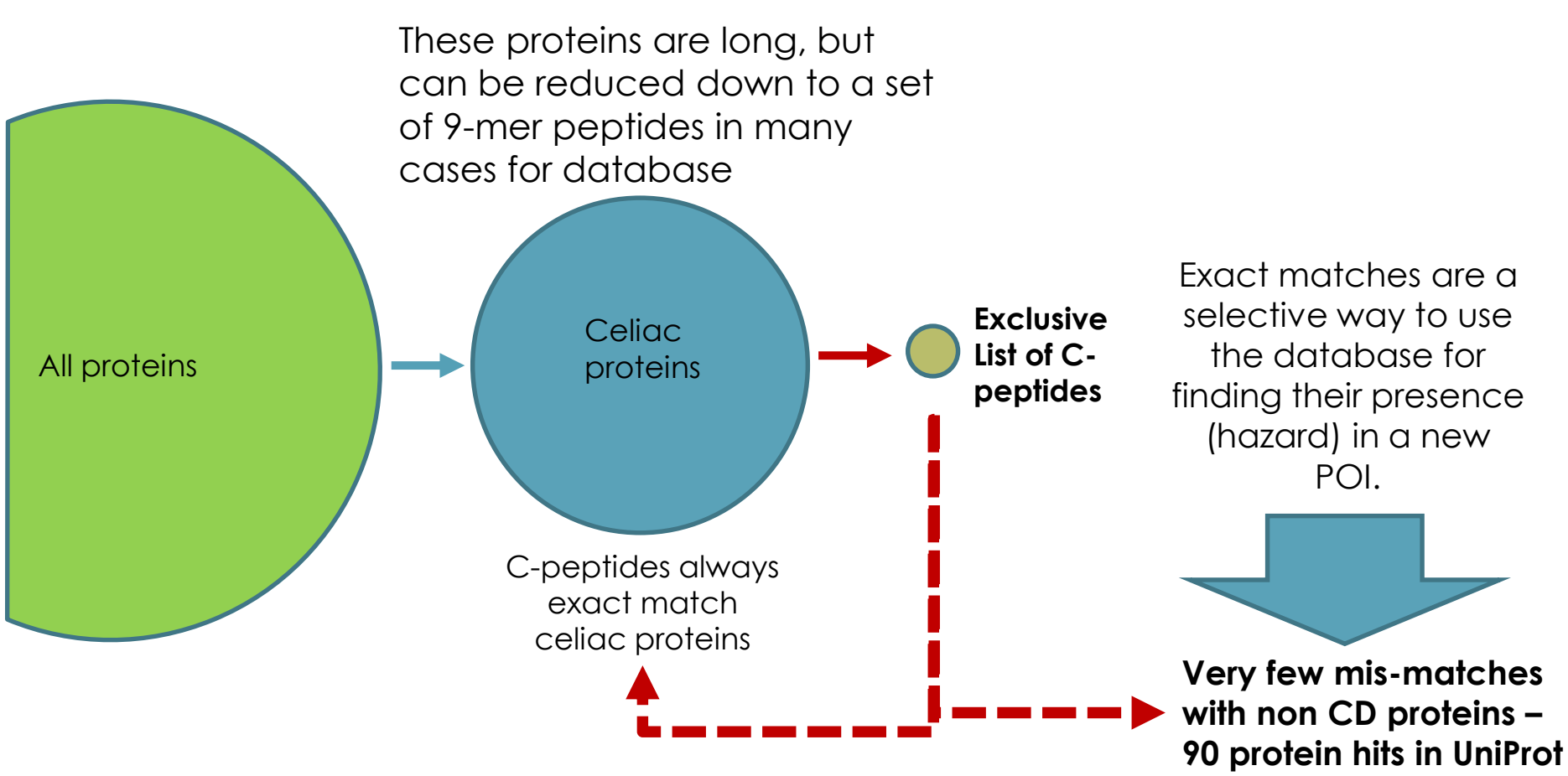
Defining a Celiac Sequence Database

- Multiple resources
  - The 26 peptides from Sollid et al., 2012
  - The 1,013 peptides from Allergenonline.org (Nebraska)
  - The 334 peptides from ProPepper™ (website)

- The collection of all three sources have varying lengths of peptides and have a **clear overlap** in membership (i.e., ProPepper overlaps with FARRP).

- There is evidence that the combination of the bioinformatics methods would **benefit from a reduced set of peptides**

Exact Matches work well – Works with the large database as well as with anyone of the three sources



Extending past Exact Matches - motifs and partial matches. are negatively impacted by larger celiac database and CD peptides that don't contain motifs

SQPQQ**QFPQ**
PQPQQ**QFPQ**
QQP**QQPF**PQ
PQP**QQPF**CQ
QQPFP**QQPQ**
QQP**QQPF**PQ
QQP**QQPY**PQ

Motif screen

What's in a Motif ?

Database of C-Peptides

Original intent of FARRP – exact matching[1]

9mer Imperfect Matches

9mer Exact Matches

Provides accuracy
Provides sensitivity

Little accuracy – estimate a 46-54% accuracy rate in blind screening

Motif

High confidence in outcome

No known accuracy – the partial matching is a derivative of the motif, but without good registration to known binding

Partial

Allowing **3 mismatches** changes from having a set of known exact matches to unexplainable matches @ ~**8,000 false positives**

1- Personal communications, Copenhagen, 2018

Concluding Points

Database

- Three sources available
- Specific to glutenins and gliadins when verified against UniProt by **exact matching**.

Exact Matching – This is the most useful approach for testing unknown GM proteins

- Accurate and specific for the known celiac proteins
- Partial matches can be "covered" by exact matches and thus, "validated".

Motif Matching – Only the relevant motifs offer utility in matching

- The degenerate, all combinations approach allows irrelevant matches to be obtained. **Not all motifs in any one of the three database sources.**
- There is no evidence that 4-AA motif matches act to predict the presence of a CD 9-mer celiac peptide better than a 9-mer match itself.

Partial Matching

- Derivative – its based on understanding one of the **50** motifs defined by the degenerate motif.
- **There is no list of "partial" matches that have been validated for HLA and/or T-cell binding.**

Use of a more targeted database and the relevant 4-AA motifs would retain accuracy (specificity & sensitivity). Otherwise **high proportion of false positivity** because of short length