A History of Bioinformatics: Development of in silico Approaches to Evaluate Food Proteins

Andre Silvanovich Ph. D.
Bayer Crop Sciences
Chesterfield, MO
October 2018

Bioinformatic Evaluation is a Key Component in the Hazard Assessment of Food Proteins

Transgenic crops typically contain genes/proteins that confer a desirable trait(s) such as:

- Insect and virus resistance
- Herbicide tolerance
- Nutritional enhancement
- Improved storage stability

A hazard assessment of introduced protein(s) is one of many evaluations conducted and submitted to regulators prior to product approval

Bioinformatics establishes a framework that along with *in vivo* and *in vitro* testing characterize food proteins

- It tells us what a protein is and if it is novel or a homolog of an existing food protein
- Likewise through knowledge of identity, we can state the converse

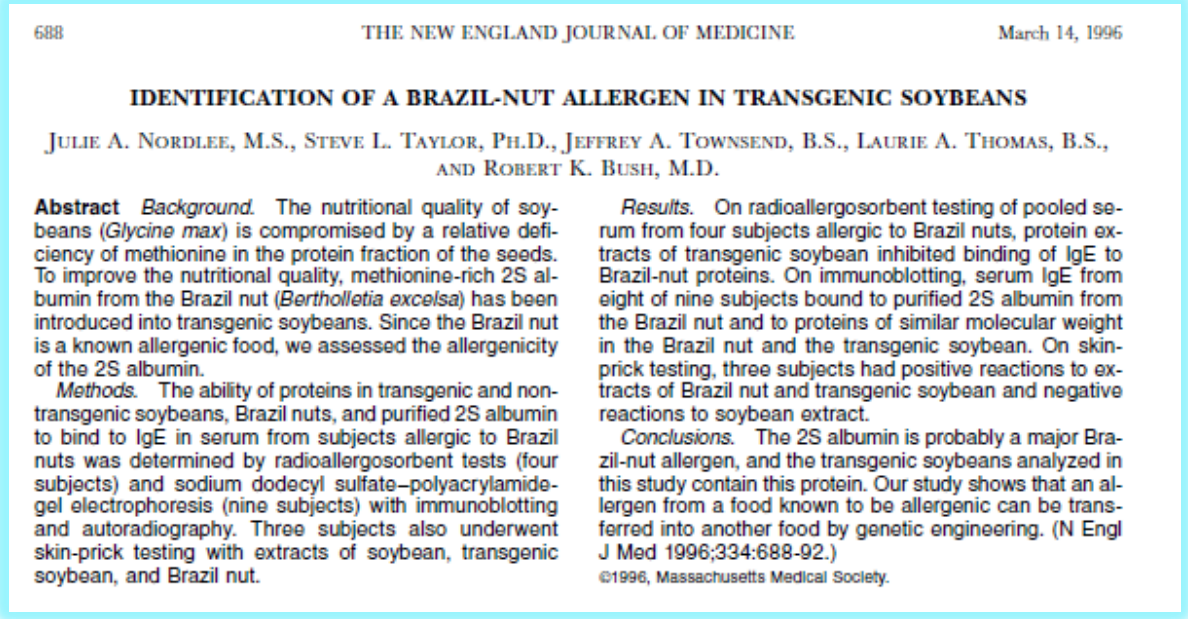
Specific Bioinformatic Applications

Protein prospecting

- Identify genes/proteins that may confer a desirable trait
- Assess allergenic and toxic potential of candidate proteins

Protein hazard identification

- Source organism
- Identity/function/mode of action
- Potential allergenicity
- Potential toxicity or other undesirable characteristic



688 THE NEW ENGLAND JOURNAL OF MEDICINE March 14, 1996

IDENTIFICATION OF A BRAZIL-NUT ALLERGEN IN TRANSGENIC SOYBEANS

JULIE A. NORDLEE, M.S., STEVE L. TAYLOR, PH.D., JEFFREY A. TOWNSEND, B.S., LAURIE A. THOMAS, B.S., AND ROBERT K. BUSH, M.D.

Abstract *Background.* The nutritional quality of soybeans (*Glycine max*) is compromised by a relative deficiency of methionine in the protein fraction of the seeds. To improve the nutritional quality, methionine-rich 2S albumin from the Brazil nut (*Bertholletia excelsa*) has been introduced into transgenic soybeans. Since the Brazil nut is a known allergenic food, we assessed the allergenicity of the 2S albumin.

Methods. The ability of proteins in transgenic and nontransgenic soybeans, Brazil nuts, and purified 2S albumin to bind to IgE in serum from subjects allergic to Brazil nuts was determined by radioallergosorbent tests (four subjects) and sodium dodecyl sulfate–polyacrylamide-gel electrophoresis (nine subjects) with immunoblotting and autoradiography. Three subjects also underwent skin-prick testing with extracts of soybean, transgenic soybean, and Brazil nut.

Results. On radioallergosorbent testing of pooled serum from four subjects allergic to Brazil nuts, protein extracts of transgenic soybean inhibited binding of IgE to Brazil-nut proteins. On immunoblotting, serum IgE from eight of nine subjects bound to purified 2S albumin from the Brazil nut and to proteins of similar molecular weight in the Brazil nut and the transgenic soybean. On skin-prick testing, three subjects had positive reactions to extracts of Brazil nut and transgenic soybean and negative reactions to soybean extract.

Conclusions. The 2S albumin is probably a major Brazil-nut allergen, and the transgenic soybeans analyzed in this study contain this protein. Our study shows that an allergen from a food known to be allergenic can be transferred into another food by genetic engineering. (N Engl J Med 1996;334:688-92.)

The value of bioinformatic hazard identification traces its origin to Nordlee et al., 1996

- Nutritionally enhanced soybean
- Increase methionine content with Brazil Nut 2S albumin
- Brazil Nut 2S albumin allergenic
- Brazil Nut allergic individuals react to soybean extracts containing Brazil Nut 2S albumin transgene

Bioinformatic Methods

Comparison of novel protein sequence(s) with databases

The bioinformatic process is relatively unchanged over last 20+ years (query, software tool, database search)

Fit for purpose

- Is the novel protein a homolog of a known allergen or toxin?

Sequence aligners:

- FASTA or BLAST

Databases:

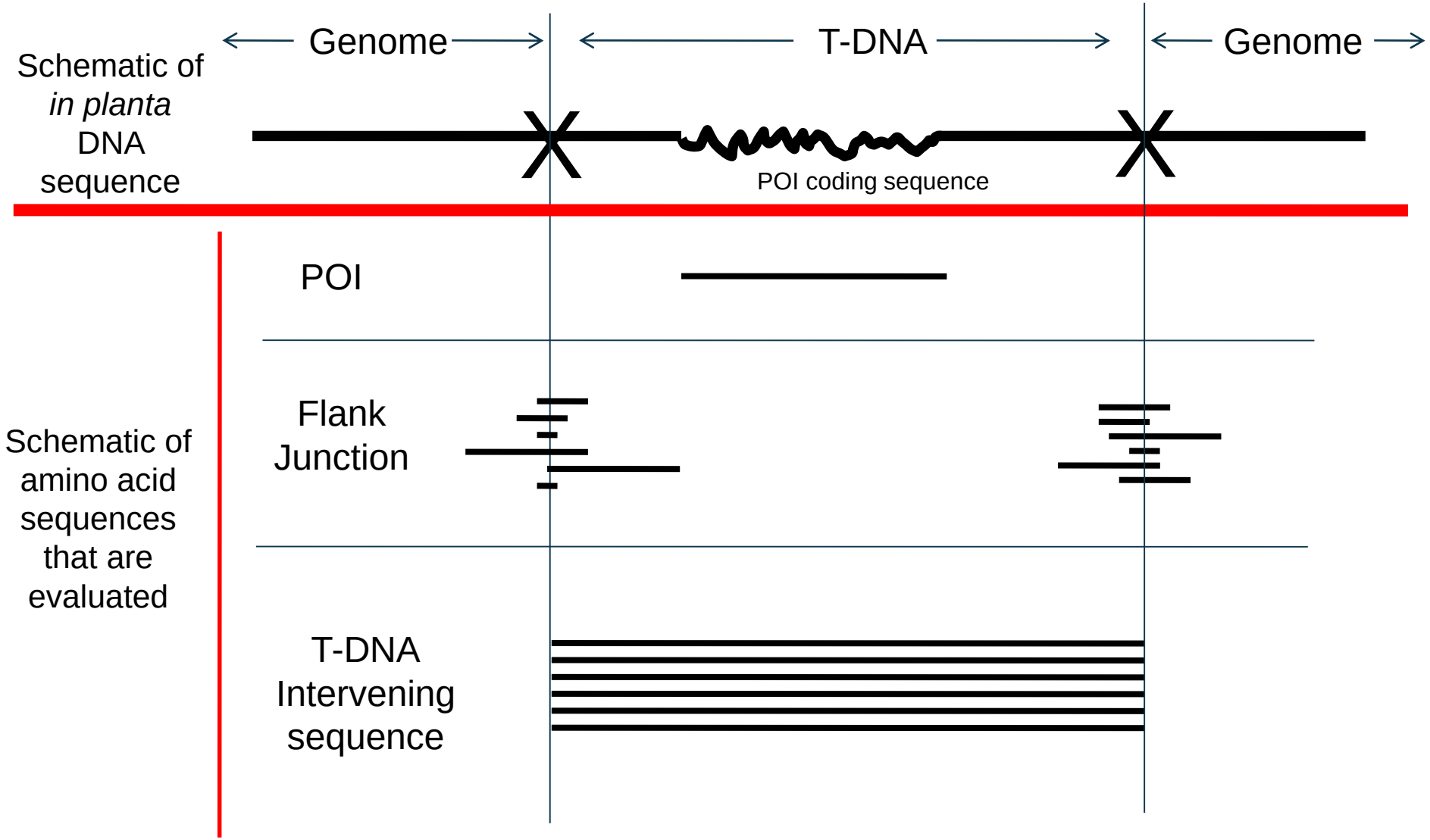
- Comprehensive protein
- Curated subsets: Allergen and Toxin

Interpretation of alignment data

- Predefined thresholds
  - Expectation value
  - % identity in window
- Inspection of alignments, identification of domains

Draw conclusions of hazard/toxicity/allergenic potential

Query Sequences



Evolution Bioinformatic Methods

Databases

Sliding window short mer searches to assess allergenicity

Sliding window aligner searches to assess allergenicity

Sliding window short mer searches to assess celiac potential

E-values

The influence of high throughput sequencing and automated annotation

Databases

Allergen and toxin homolog identifications may benefit from searching curated databases

Helps to provide a focal point for these critical analyses

- Searching comprehensive databases may yield thousands of alignments

Through curation high levels of stringency can be applied to the database

Curated allergen databases first came into use in the late 1990's

- Evolved over the years
  - Assembled by individual companies
  - Assembled by FARRP, AOL
  - Assembled through HESI, COMPARE
- Updated annually
- ~ 2000 proteins spanning ~220 families

Curated toxin databases pose specific challenges

- What is a "toxin" and the context for "toxicity"?

Sliding Window Short mer Searches to Identify IgE-Binding Epitopes

Mer by mer comparison between novel protein and allergen database

- Exact matches with any identical length mer from any allergen
- Modelling based upon chance correlates well with real life observations
  - Hit frequency increases with query protein length
- Window size 6 or 8 amino acids
  - 6-mer ~75% of all proteins would yield at least one match
  - 8-mer ~13% of all proteins would yield at least one match
- >99% false positive rate, most mers from known allergens are not responsible for IgE-binding
  - Short peptides such as Poly-G or Poly-P
- False negatives due to discontinuous epitopes
- Although value is limited, some world areas still request the search be performed

Sliding Window Aligner Searches

Threshold of 35% identity over a window of 80 aa

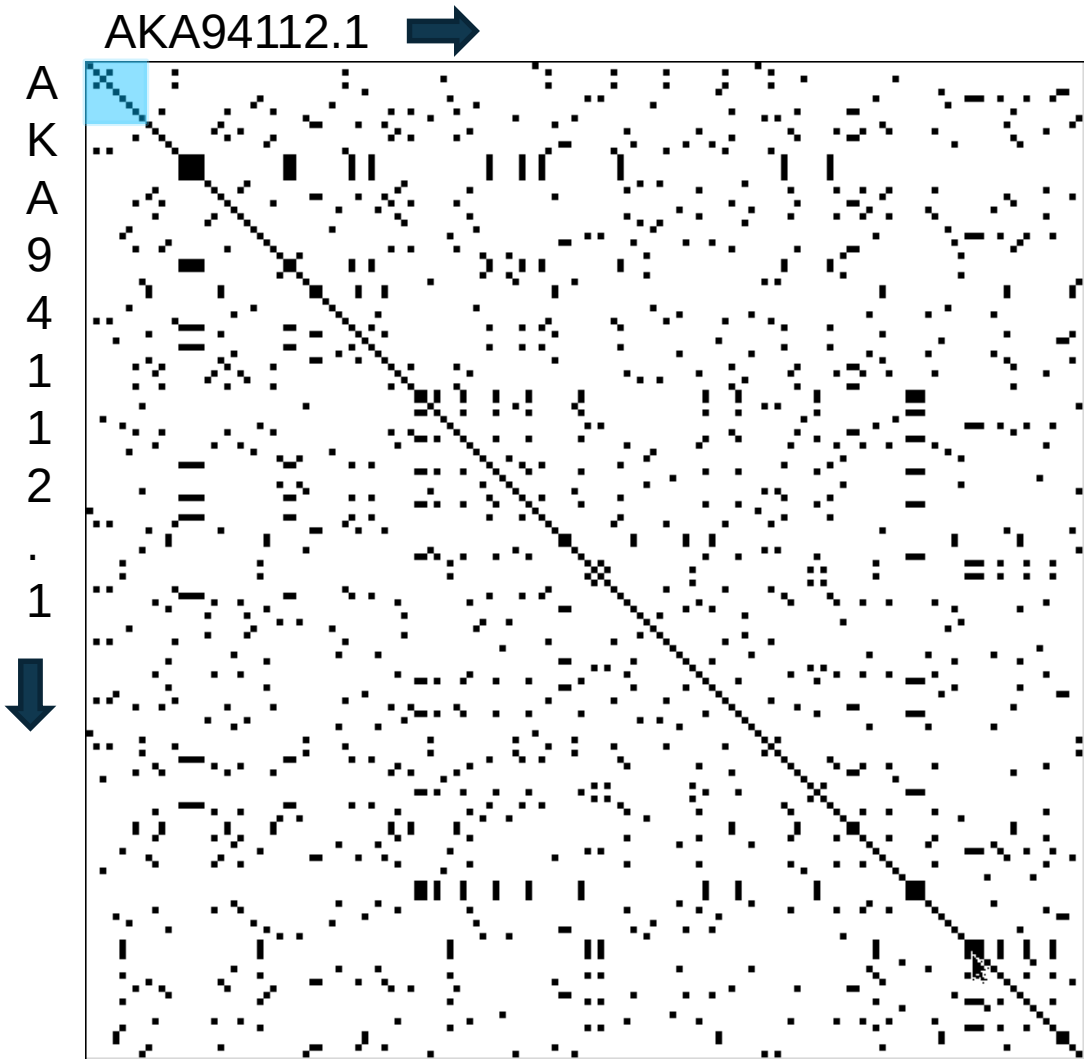
Based upon a single exceptionally well defined allergen family, *Bet v 1*

80 aa is thought to reflect the size of a “typical domain”

Subject increased false positive rate vs. intact query sequene due to optimization

- Optimization is the insertion of gaps into sequences to maximize alignment length
- Gap insertion at beginning and end of sequence is less costly than the middle
- Sliding window alignments displays an edge effect
- The sliding window is removed from the context of the full protein

Exact Match in a Diagonal Matrix for Rubisco Fragment




```
>AKA94112.1 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
WRDRFLYVMEGVNRAAAASGEVKGSYLNVTAATMEECYKRAEFAKEVGSVIIMIDLVIGYTAIQTMAIWARENNMI
LHLHRAGNSTYSRQKNHGINFRVISKWMRMAGVDHIHAGTVVGKLEGDPIIIKGFYNTLLLPKLEVNLPQGLFFEM

>AKA94112.1 ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
WRDRFLYVMEGVNRAAAASGEVKGSYLNVTAATMEECYKRAEFAKEVGSVIIMIDLVIGYTAIQTMAIWARENNMI
LHLHRAGNSTYSRQKNHGINFRVISKWMRMAGVDHIHAGTVVGKLEGDPIIIKGFYNTLLLPKLEVNLPQGLFFEM
```

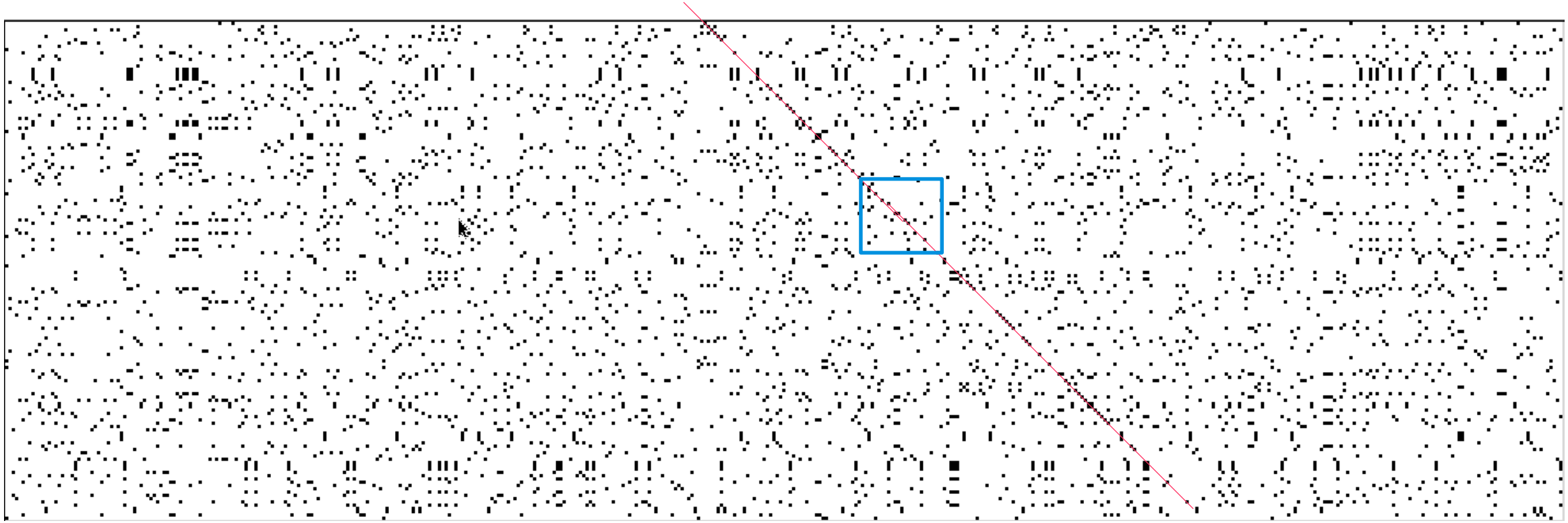
All sequence aligning tools employ the concept of a diagonal matrix

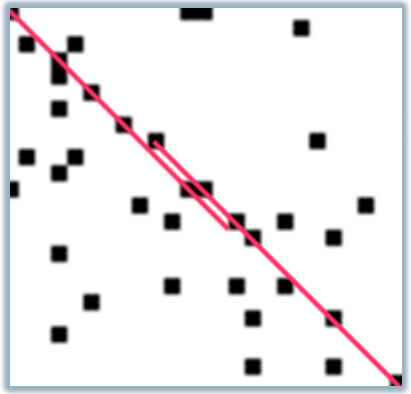
When a position by position comparison is made between two identical sequences, a diagonal is produced on a matrix

	W	R	D	R	F	L	Y	V	M
W	X		X						
R		X							
D	X		X						
R				X					
F					X				
L						X			
Y							X		
V								X	
M									X

Regardless of the length of a query and database sequence, each search is done position by position

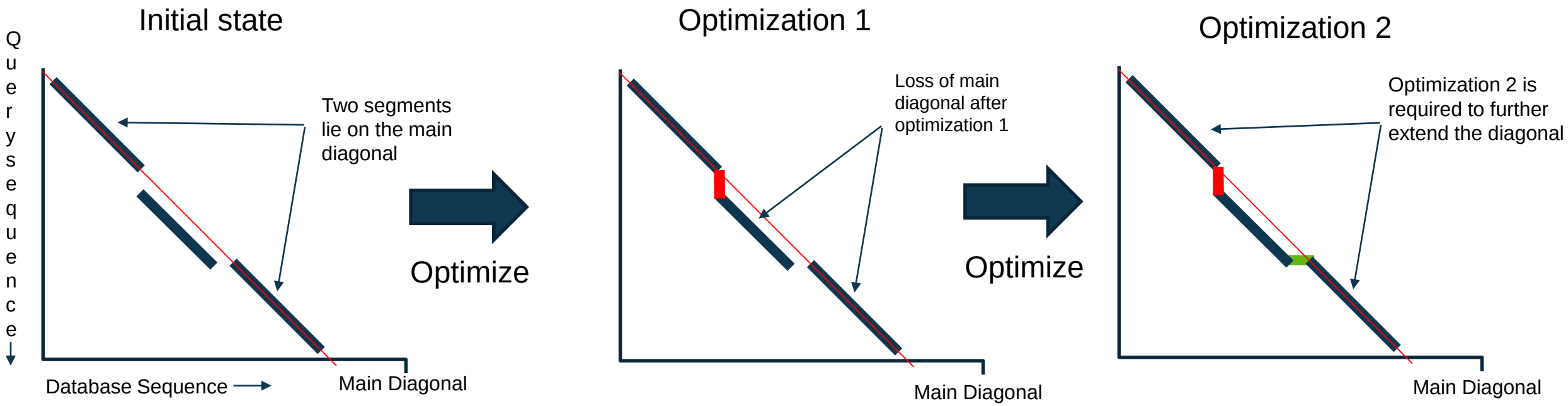
Diagonal Matrix of Rubisco Homologs





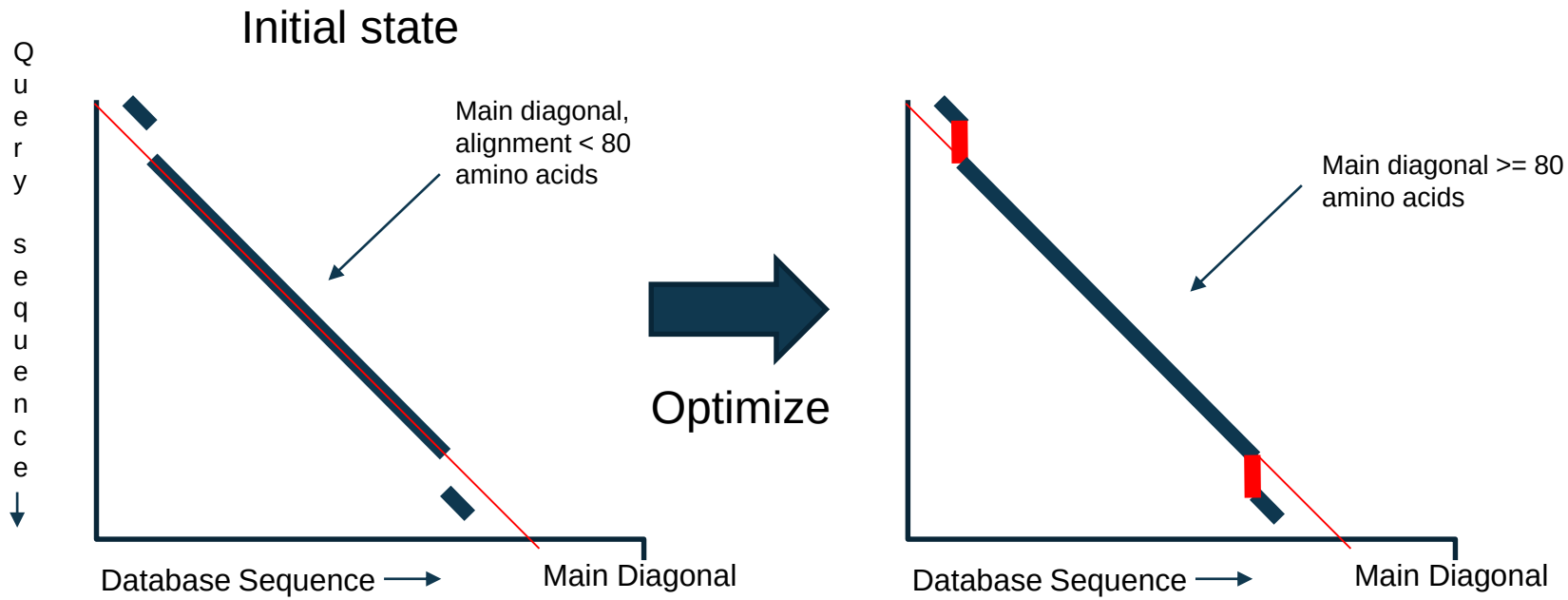
- Homologs will often display discontinuities in diagonals.

Optimization is the Process of Building the Longest Possible Diagonal



- There is a negative cost associated with adding the gap
- The cost of the gap must be outweighed by the added alignment length/amino acid composition
- Optimizations in the middle of an alignment frequently create the need for an additional optimization

Gaps Near Ends of a Sequence



- Optimization at ends does not require a second compensating optimization (edge effect)
- Since optimizations at ends are less restricted, they occur more frequently and are more likely to be influenced by high scoring amino acids
- An intact sequence has 2 ends
- Sliding windows introduce hundreds of ends, each end permitting optimization that is out of the context of the complete protein sequence
- In extreme cases such as glutamine-rich sequences, an 80 amino acids query can yield an alignment of 110 amino acids in length

Sliding 80 Amino Acid Window Search Summary

- Several publications compare 80 aa sliding window vs. full length query sequences using *E*-value threshold for homolog identification
- *E*-values that display comparable sensitivity but with lower off target false positive rates have been proposed
- Since similarity is not considered, some homologs that display high levels of similarity will be missed by an identity only threshold
- Homologs are best identified across the entire length by including identity and similarity, once identified domains can then be identified/assessed

Sliding Window Short mer Searches to Assess Celiac Potential

- Alignment with degenerate 4 aa peptide sequence
- 9-mer comparison with library of celiac disease implicated peptides
  - Up to 3 mismatches permitted
- High match frequencies observed
  - 7-10% of all proteins are flagged by the degenerate 4 aa peptide
- Does not include peptide context within protein
- More information on this topic Thursday

The Influence of High Throughput Sequencing and Automated Annotation

Vast increase in publicly available sequence data due to High Throughput Sequencing

Protein sequence data is the result of gene prediction and auto annotation

- No physicochemical analysis of auto annotated sequence
- Little, if any, human oversight or QC review submitted
  - Some auto annotation is in error
  - Some auto annotation could be more precise
- Creates issues for custom datasets such as an allergen or toxin database
  - Numbers of auto annotated sequences preclude human review
  - Annotation issues tend to be inconsistently propagated
  - Identification of erroneous auto annotations is labor intensive

The Influence of High Throughput Sequencing and Automated Annotation cont..

“In a study published Tuesday in PLOS Biology, researchers at Northwestern University reported that of our 20,000 protein-coding genes, about 5,400 have never been the subject of a single dedicated paper.”

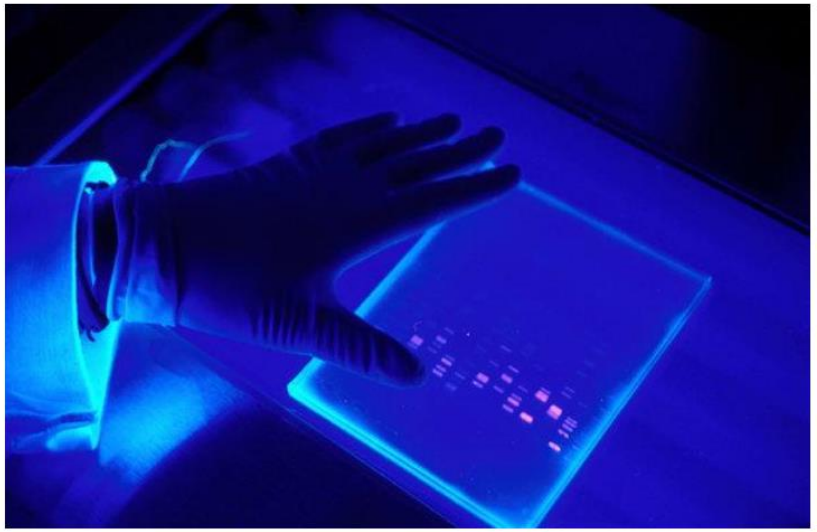
“Most of our other genes have been almost as badly neglected, the subjects of minor investigation at best. A tiny fraction — 2,000 of them — have hogged most of the attention, the focus of 90 percent of the scientific studies published in recent years.”

Curating sequences for custom databases provides a means of establishing quality criteria

Alignment data must be reviewed by skilled scientists to draw meaningful conclusions

Why Your DNA Is Still Uncharted Territory

Scientists are focusing on a relatively small number of human genes and neglecting thousands of others. The reasons have more to do with professional survival than genetics.



DNA being sequenced in a process called gel electrophoresis. An enormous chunk of the human genome remains unstudied despite significant technological advances. Eurelios/Science Source

By Carl Zimmer

Sept. 18, 2018

14

Cited text viewed and image downloaded Sept.19, 2018

https://www.nytimes.com/2018/09/18/science/why-your-dna-is-still-uncharted-territory.html?action=click&module=Discovery&pgtype=Homepage

Summary

The bioinformatic tools used for homolog identification are fit for purpose

- Homolog identification is then used in hazard assessment

Arbitrary window length searches are prone to high false positive identification rates

- False positive rates can be decreased with increased stringency
  - Additional downstream criteria
  - Look for peptide windows within alignments created using full length queries

Validation of alignment thresholds and use of software tools as developers intended would be beneficial

- Proteins differ from one another based upon physicochemical characteristics
- Likewise, bioinformatic homolog identifications benefit from flexibility

High Throughput Sequencing will continue to inform and complicate interpretation of bioinformatic data