Biological Chemistry Laboratory
Biology 3515/Chemistry 3515
Spring 2022

Lecture 8:
More on Curve Fitting

and

Introduction to Proteases

Thursday, 3 February 2022

University of Utah
goldenberg@biology.utah.edu

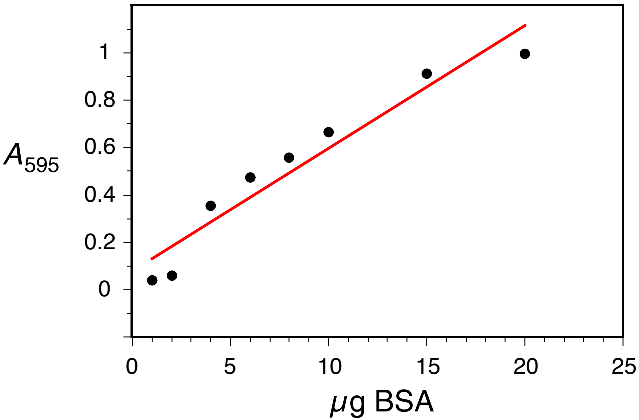
Computer Labs

- Computer Labs this week and next.
  - Start at 2:00 PM!
  - Room 150 Biology Building
- This week: Graphing and curve fitting with SciDAVis.
- Next week: Molecular modeling with PyMOL.
- We will use the computers in the lab, not personal laptops.
- But, you should still install SciDAVis and PyMOL on your own computer. Use the versions available on Canvas.

First Quiz: Thursday, 10 February

- In class, during second half.
- Study materials:
  - Quizzes from previous years (Canvas).
  - Problems in lab manual.
  - Answers will not be posted, but the TAs and instructors are available for discussion.
- Review session to be scheduled for next Wednesday.

How Do We Judge the Goodness of Fit?

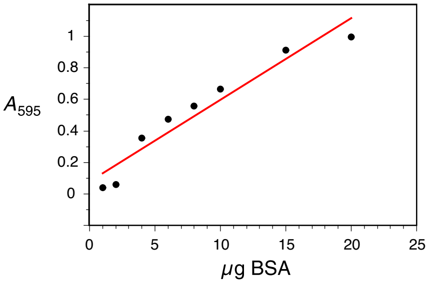


- The final, minimized, value of χ^2 for this fit:

$$\chi^2 = 0.062$$

- Can use this value to compare this fit to fits of other functions to the same data.
- But, actual value does not have a clear meaning.
- Adding more measurements will almost always increase χ^2.

The Coefficient of Determination, R^2



- What fraction of the total variation of y-values is accounted for by the fit function?
- For this fit:

$$R^2 = 0.93$$

Compare χ^2 to the total variation in y-values.

- Define the total sum of squares:

$$SS_{\text{tot}} = \sum (y_i - \bar{y})^2$$

$\bar{y} =$ mean of y-values.

- The ratio:

$$\frac{\chi^2}{SS_{\text{tot}}}$$

represents the fraction of the variation that is *not* accounted for by the fit function.

- The fraction of variation that *is* accounted for the function:

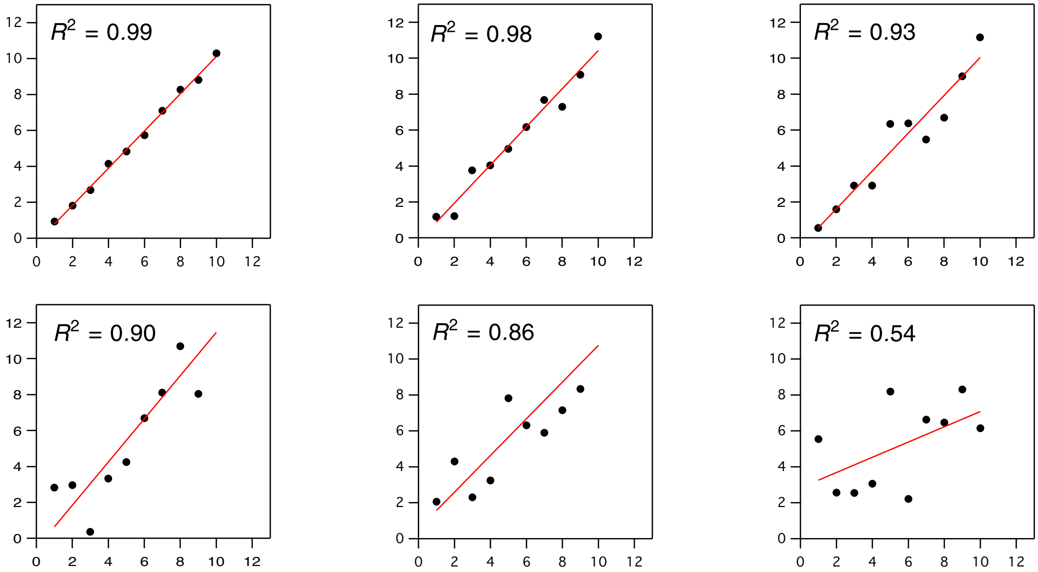
$$R^2 = 1 - \frac{\chi^2}{SS_{\text{tot}}}$$

R^2 should lie between 0 and 1.

Interpreting R^2

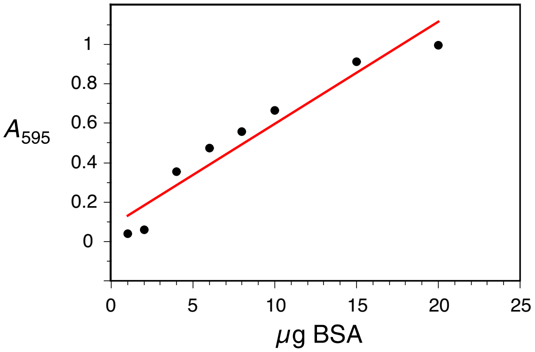
- $0 \leq R^2 \leq 1$
- $R^2 = 0$: No correlation between x and y in the experimental data.
- $R^2 = 1$: A "perfect" fit of the experimental data to the function.
- R^2 is the fraction of the total variation of the experimental y values that is accounted for by the linear function.
- R^2 is often called the *correlation coefficient*, which is a related but different parameter.

Some Examples



Clicker Question #1

What if the fit isn't as good as we'd like?



$R^2 = 0.93$

Should we:

A) Delete some points?

B) Find a function that better represents the data?

C) Accept that there is some error in our measurements?

D) Repeat the experiment more carefully?

All answers count (for now)!

Polynomials as Fitting Functions

- General form of a polynomial function:

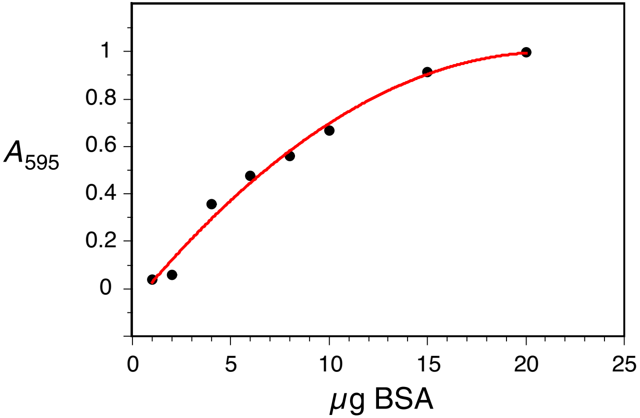
$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$

- A polynomial in which the largest power of x is x^n is called an n^{th}-order polynomial.
    - A first-order polynomial is a straight line: $y = a_0 + a_1 x$
    - A second-order polynomial is also called a quadratic function:

      $y = a_0 + a_1 x + a_2 x^2$
    - A third-order polynomial is also called a cubic function:

      $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$
- An n^{th}-order polynomial contains $n + 1$ coefficients $(a_0, a_1, a_2, \ldots, a_n)$.
- A **minimum** of $n + 1$ data points are required to fit an n^{th}-order polynomial.

A 2nd-order Polynomial Least-squares Fit to Bradford Calibration Data



- For 2nd-order polynomial fit:

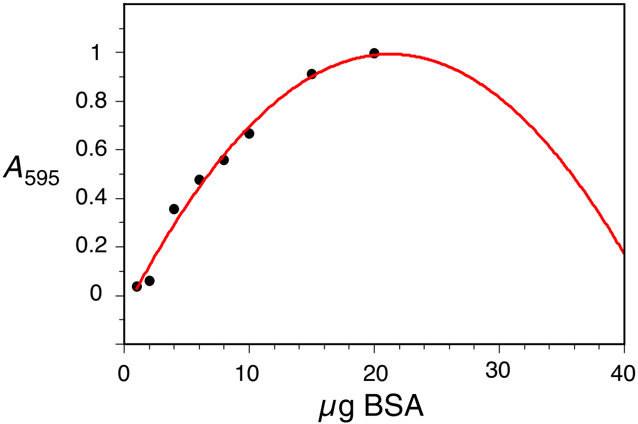
  $$\chi^2 = 0.012$$
  $$R^2 = 0.988$$

- For linear fit:

  $$\chi^2 = 0.062$$
  $$R^2 = 0.93$$

- Increasing the number of parameters almost always improves the fit!
- Is it justified here?

Does the Fit Function Make Sense Physically?

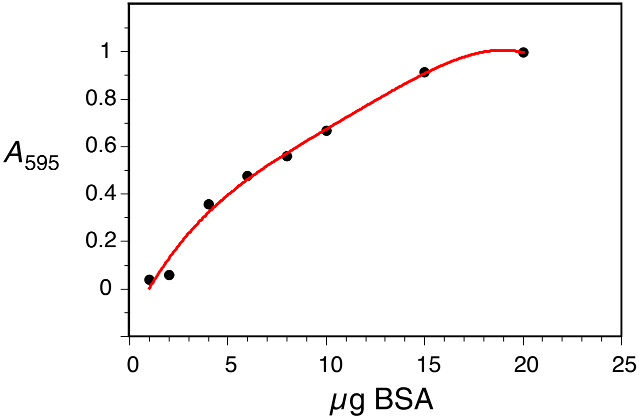


- Should the absorbance decrease as the amount of BSA increases beyond 20 μg?

  Probably not!

- The function could serve as a calibration curve over the range used to fit it, but not beyond.

A 4^{th}-order Polynomial Least-squares Fit to Bradford Calibration Data



- For 4^{th}-order polynomial fit:

$$\chi^2 = 0.01$$
$$R^2 = 0.991$$

- For 2^{nd}-order polynomial fit:

$$\chi^2 = 0.012$$
$$R^2 = 0.988$$

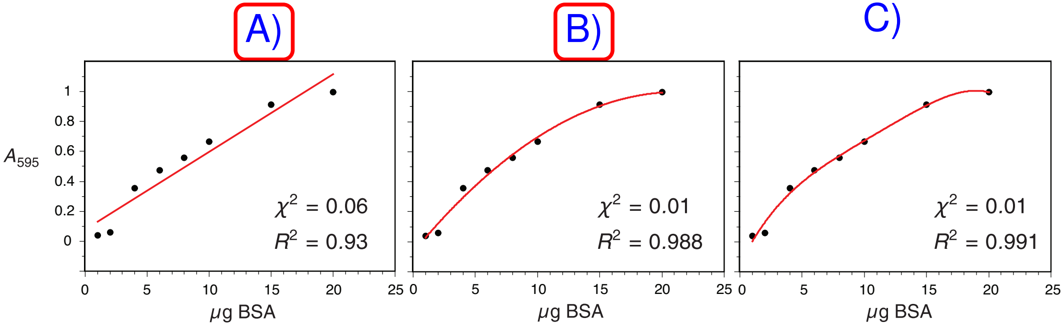
- For linear fit:

$$\chi^2 = 0.062$$
$$R^2 = 0.93$$

- Have we gone to far?

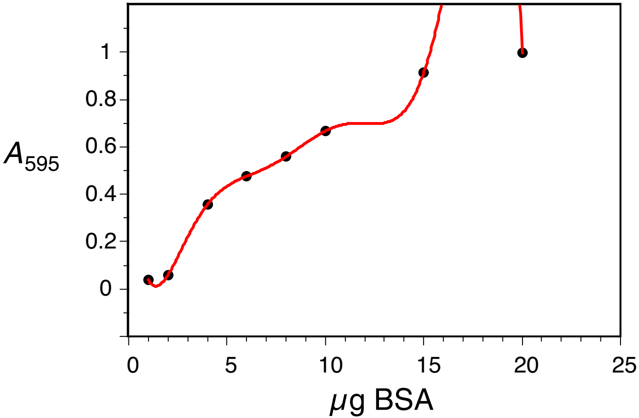
Clicker Question #2

Which is the most reasonable fit?



All answers count (for now)!

A 7th-order Polynomial Least-squares Fit to Bradford Calibration Data with 8 Points



- For 7th-order polynomial fit:

  $$\chi^2 = 0$$
  $$R^2 = 1$$

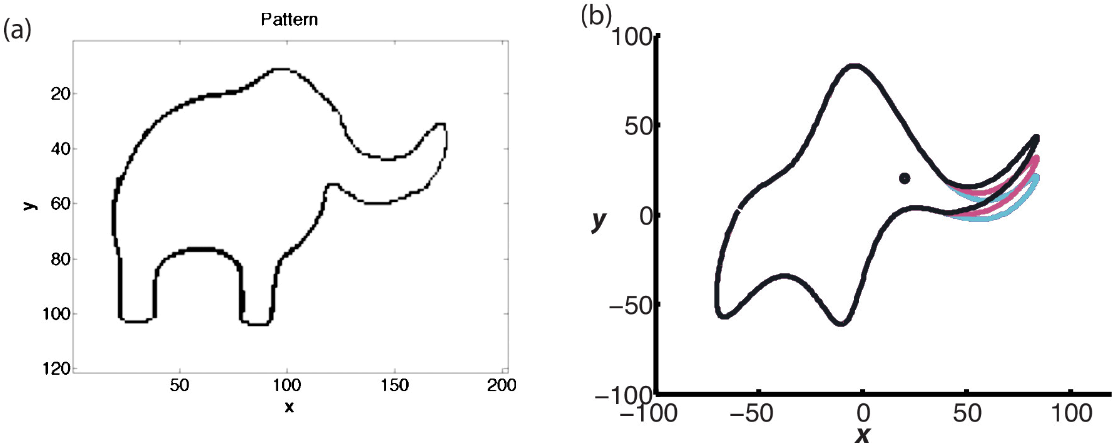
  A perfect fit!

  Or, perfectly absurd?

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

John von Neumann, according to Enrico Fermi, as quoted by Freeman Dyson. *Nature* (2004) **427**, 297

Fitting an Elephant



Mayer, J., Khairy, K. & Howard, J. (2010). Drawing an elephant with four complex parameters. *Am. J. Phys.*, 78, 648–649.
http://dx.doi.org/10.1119/1.3254017

Another Interesting Function

$$y = \frac{ax}{b + x}$$

- When $x \ll b$

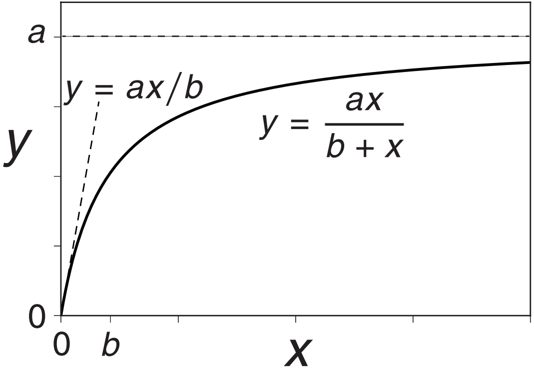
$$y = \frac{ax}{b + x} \approx \frac{ax}{b}$$

A line through the point $(0, 0)$, with slope a/b.

- When $x \gg b$

$$y = \frac{ax}{b + x} \approx \frac{ax}{x} = a$$

A constant, a.



“Linear” versus “Non-linear” Curve Fitting

- In the context of curve-fitting, a polynomial

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$$

  is said to be a “linear” function in the sense that y is a linear function of each of the fit parameters, a_i (even if it isn't linear with respect to x).

- Equations of this type can be fit to data relatively easily using equations like those shown for the straight line fit.

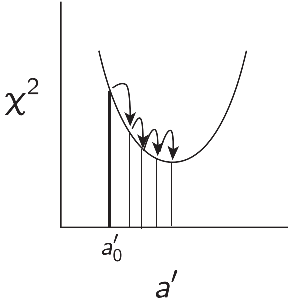
- The equation for a rectangular hyperbola:

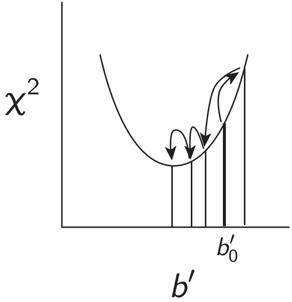
$$y = \frac{a \cdot x}{b + x}$$

  is *not* linear with respect to the parameter b.

- For non-linear equations, least-squares fitting usually must be done iteratively.

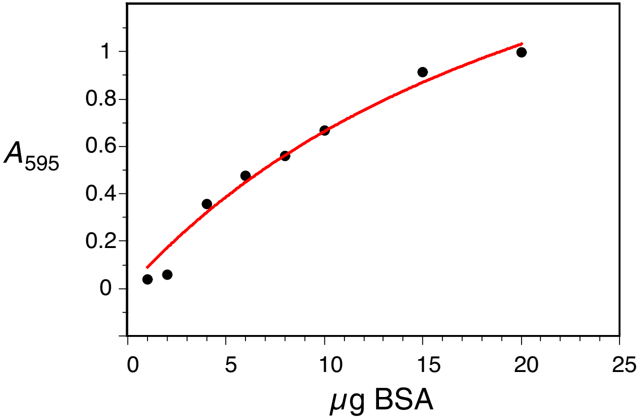
An Iterative Method to Minimize χ^2





1. Make initial estimates of parameters a and b
2. Calculate χ^2
3. Change the parameters a little bit and recalculate χ^2
4. If χ^2 decreases, change the parameters some more in the same direction; otherwise change, the parameters in the opposite direction.
5. Repeat until χ^2 does not decrease further.

A Rectangular Hyperbola Fit to Bradford Calibration Data



- For fit to rectangular hyperbola:

  $$\chi^2 = 0.02$$
  $$R^2 = 0.977$$

  With only two parameters!

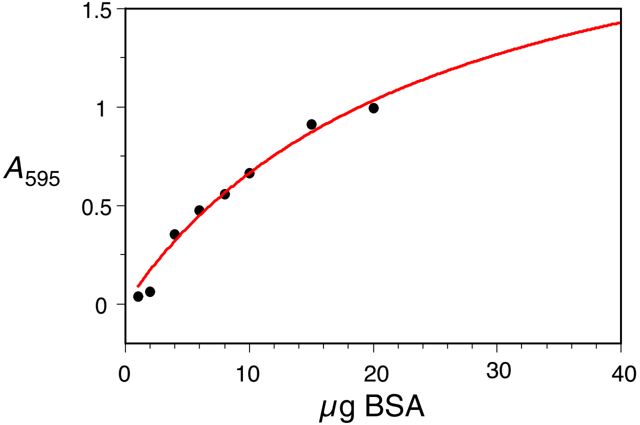
- For 2nd-order polynomial fit:

  $$\chi^2 = 0.01$$
  $$R^2 = 0.988$$

- For linear fit:

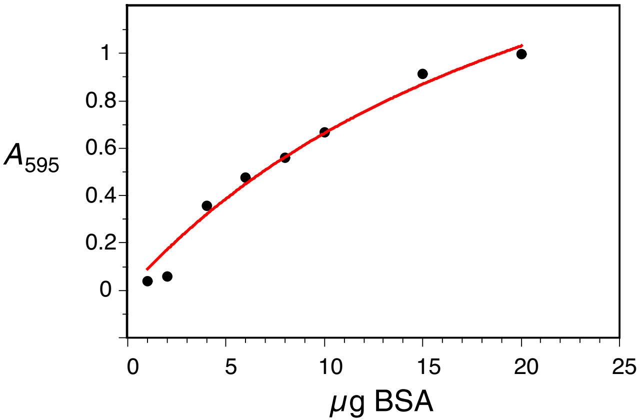
  $$\chi^2 = 0.062$$
  $$R^2 = 0.93$$

Does the Fit Function Make Sense Physically?



- Does the extrapolation look plausible?
- Is the curvature real?
- How could we find out?
- Why might the Bradford calibration curve have this shape?

A Rectangular Hyperbola Fit to Bradford Calibration Data



- Fit function:

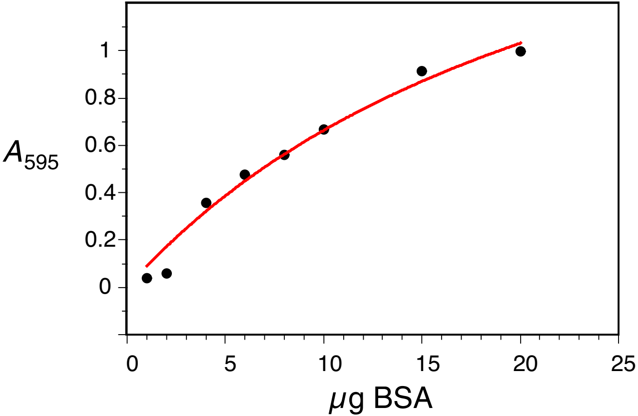
$$y = \frac{ax}{b + x}$$

- Fit parameters:

$$a = 2.32 \pm 0.53$$
$$b = 24.9 \pm 6.6$$

- What are the units for the parameters?

A Rectangular Hyperbola Fit to Bradford Calibration Data



- Fit function:

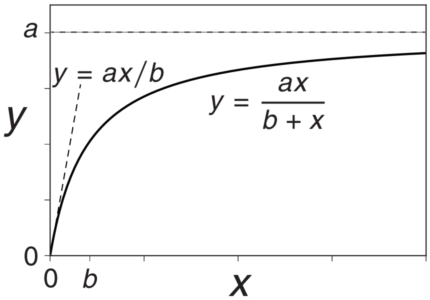
$$y = \frac{ax}{b + x}$$

- Fit parameters:

$$a = 2.32 \pm 0.53$$
$$b = 24.9 \pm 6.6$$

- Why are the uncertainties so large?

Why Are the Uncertainties So Large?



- When x is small relative to b:

  $$y = \frac{ax}{b+x} \approx \frac{ax}{b}$$

  A line through the point $(0, 0)$, with slope a/b.

  If we only have data in this region, the slope, a/b, is well defined, but lots of pairs of a and b will fit the data well.

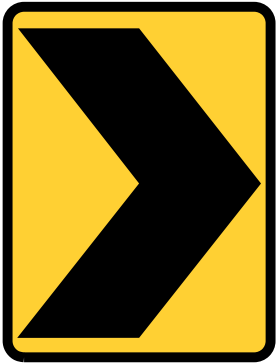
- When x is large relative to b:

  $$y = \frac{ax}{b+x} \approx \frac{ax}{x} = a$$

  If we only have data in this region, what will happen to our fit?

- To determine both a and b, we need data over a range that includes values that are less than b and values that are greater than b.
- Good data analysis requires good experimental design! (And, good data!)

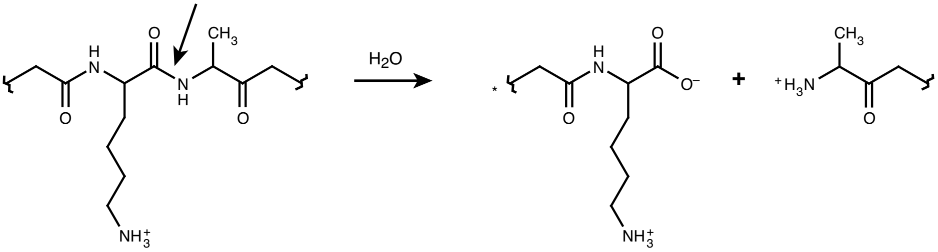
Warning!



Direction Change

Introduction to Proteases

The General Protease Reaction

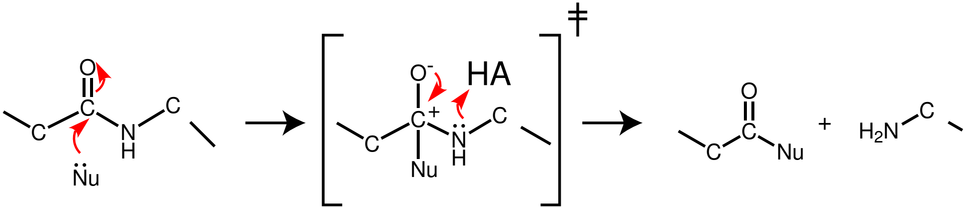


- About 2% of genes in most organisms encode proteases. (Hedstrom, L. 2002, *Chem. Rev.* **102**, 4429)

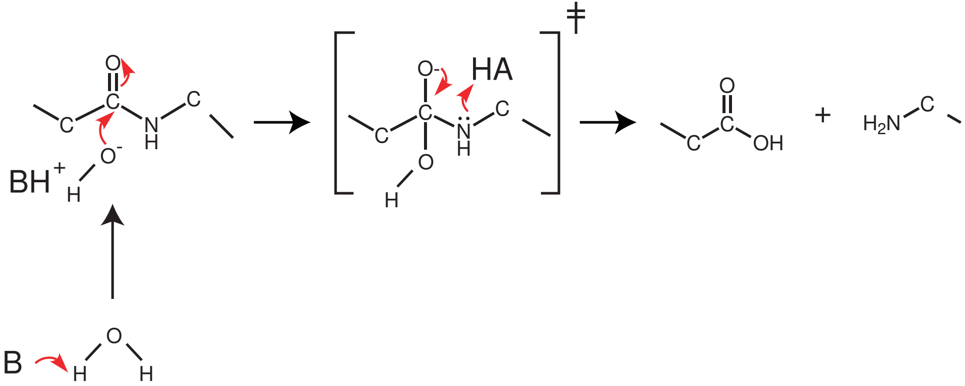
Some Biological Functions of Proteases

- Digestion of food
  - Not very selective
  - Catalyzed by trypsin, chymotrypsin, pepsin and other proteases
- Intracellular protein degradation
  - Highly selective and regulated
  - Often catalyzed by large protein complexes, *e.g.,* the proteasome
- Regulation of biological activity by proteolytic activation
  - Angiotensin converting enzyme (blood pressure regulation)
  - Blood clotting and disruption of blood clots
  - Complement fixation (an element of the immune response)
  - Apoptosis (programmed cell death)
- Maturation of viral proteins, *e.g.,* HIV, coronaviruses and many others

General Protease Mechanism is Nucleophilic Substitution



Water Can Act as the Nucleophile, but Must be Activated by a Base



- Why is this reaction so slow in the absence of an enzyme?
- How do enzymes enhance the rate?