Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Ligand binding curves can be fitted by one binding site models or multiple binding sites with different affinity. A distinction has to be made between high/low affinity receptor sites that are independent and can only show negative cooperativity, and allosteric and other site-site interactions that can also give positive cooperativity.

Example 1: Ligand varied mode

From the main SIMFIT menu select [A/Z], open program **hlfit**, select the ligand-varied option, and view the default test file `hlfit.tf4` which has the following data.

x	y	$se(y)$
0.021759	0.19832	0.0091144
0.021759	0.19438	0.0091144
0.021759	0.18094	0.0091144
0.039440	0.30473	0.0047306
0.039440	0.29537	0.0047306
0.039440	0.29883	0.0047306
0.071490	0.46465	0.015273
0.071490	0.49460	0.015273
0.071490	0.48484	0.015273
0.12958	0.71278	0.048762
0.12958	0.67885	0.048762
0.12958	0.61663	0.048762
0.23488	0.87238	0.048295
0.23488	0.80269	0.048295
0.23488	0.89546	0.048295
0.42575	1.0246	0.044998
0.42575	1.1137	0.044998
0.42575	1.0806	0.044998
0.77172	1.4145	0.062457
0.77172	1.2934	0.062457
0.77172	1.3806	0.062457
1.3988	1.3619	0.13387
1.3988	1.6295	0.13387
1.3988	1.4897	0.13387
2.5355	1.7047	0.19446
2.5355	1.4435	0.19446
2.5355	1.8236	0.19446
4.5959	1.7486	0.043681
4.5959	1.7613	0.043681
4.5959	1.8298	0.043681

The columns contain data in the following format.

1. **Column 1**: the non–negative ligand concentration x which must be in non-decreasing order.
2. **Column 2**: the non–negative response y presumed to be dependent on fractional saturation of receptor or binding site at the concentration in column 1.
3. **Column 3**: the positive sample standard deviation of the replicate response measurements.
   This column can be omitted or set to 1 if unweighted regression is required.

To illustrate the functionality of the SimFiT program **hlfit** we shall fit a one site model followed by a two site model (or mixture of two receptor types) and see if any improvement in fit can be supported by statistical analysis. The two models are as follows.

$$f_1(x) = \frac{AK_a x}{1 + K_a x} + C$$

$$f_2(L) = \frac{A_1 K_{a_1} x}{1 + K_{a_1} x} + \frac{A_2 K_{a_2} x}{1 + K_{a_2} x} + C$$

To fit these two models, choose to start fitting at order 1 and end fitting at order 2, using the further default settings but with $C = 0$ as there is no background signal with these data. This leads to the following results tables.

Table 1: For best-fit order 1 saturation function f_1

Number	Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p	
1	A	1.7482	0.038529	1.6693	1.8271	0.0000	
2	K_a	5.2161	0.17513	4.8574	5.5749	0.0000	

Apparent Y_{max} (i.e. $A_1 + A_2 + ... + A_n$) = 1.7482
Apparent K_a (i.e. x_0 where $f(x_0) - C = Y_{max}/2$) = 0.19171

Parameter correlation matrix

1	
-0.8715	1

Table 2: For best-fit order 2 saturation function f_2

Number	Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p	
1	A_1	0.91175	0.24512	0.40790	1.4156	0.0010	
2	A_2	1.0625	0.30555	0.43439	1.6905	0.0018	
3	K_{a_1}	0.97501	0.68571	-0.43449	2.3845	0.1669	*
4	K_{a_2}	8.5829	2.0044	4.4629	12.703	0.0002	

Apparent Y_{max} (i.e. $A_1 + A_2 + ... + A_n$) = 1.9742
Apparent K_a (i.e. x_0 where $f(x_0) - C = Y_{max}/2$) = 0.31272

Parameter correlation matrix

1			
-0.9770	1		
0.9019	-0.9685	1	
0.9845	-0.9936	0.9385	1

In order to determine if a significant improvement in fit has resulted we need to consider the following questions.

1. Are the parameters well-determined with both fits ?
2. Does the residuals analysis indicate satisfactory fits ?
3. Does the F test for excess variance support model f_2 in preference to f_1 ?
4. Can the best-fit curves be seen to differ when plotted against the data ?
5. Does the graphical deconvolution display convincing evidence that both components of f_2 are contributing to the overall fit ?

The results displayed in Tables 1 and 2 show that both models fit well with parameters that differ significantly from zero. Table 3 indicates that an excellent fit has resulted for model f_2, and Table 4 supports the conclusion that there is statistical evidence that model f_2 should be accepted as explaining the data better than model f_1. This is then further emphasized by the graphical displays showing the data with best-fit curves for f_1 and f_2, and the deconvolution of the f_2 fit into the two contributing components. The concentration is often plotted on a logarithmic scale which is then proportional to chemical potential.

Table 3: Goodness of fit for model f_2

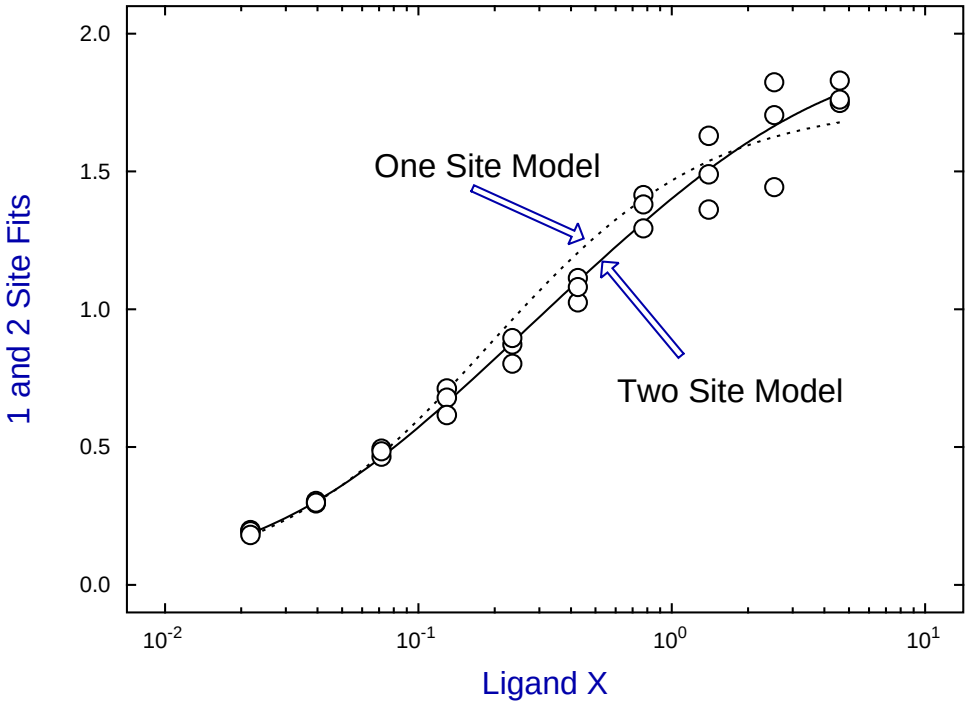
Analysis of residuals: $WSSQ$	29.952
$P(\chi^2 \geq WSSQ)$	0.2696
$R^2, cc(theory, data)^2$	0.9834
Largest Absolute relative residual	14.25%
Smallest Absolute relative residual	0.26%
Average Absolute relative residual	4.42%
Absolute relative residuals in range 0.1-0.2	6.67%
Absolute relative residuals in range 0.2-0.4	0.00%
Absolute relative residuals in range 0.4-0.8	0.00%
Absolute relative residuals > 0.8	0.00%
Number of negative residuals (m)	14
Number of positive residuals (n)	16
Number of runs observed (r)	21
P(runs $\leq r$: given m and n)	0.9820
5% lower tail point	11
1% lower tail point	9
P(runs $\leq r$: given m plus n)	0.9879
P(signs $\leq$ least number observed)	0.8555
Durbin-Watson test statistic	3.1269 > 2.5, -ve serial correlation?
Shapiro-Wilks W statistic	0.9754
Significance level of W	0.6948
Akaike AIC (Schwarz SC) stats	7.9520 (13.557)
Verdict on goodness of fit: incredible	

Table 4: F test results for model f_2 against f_1

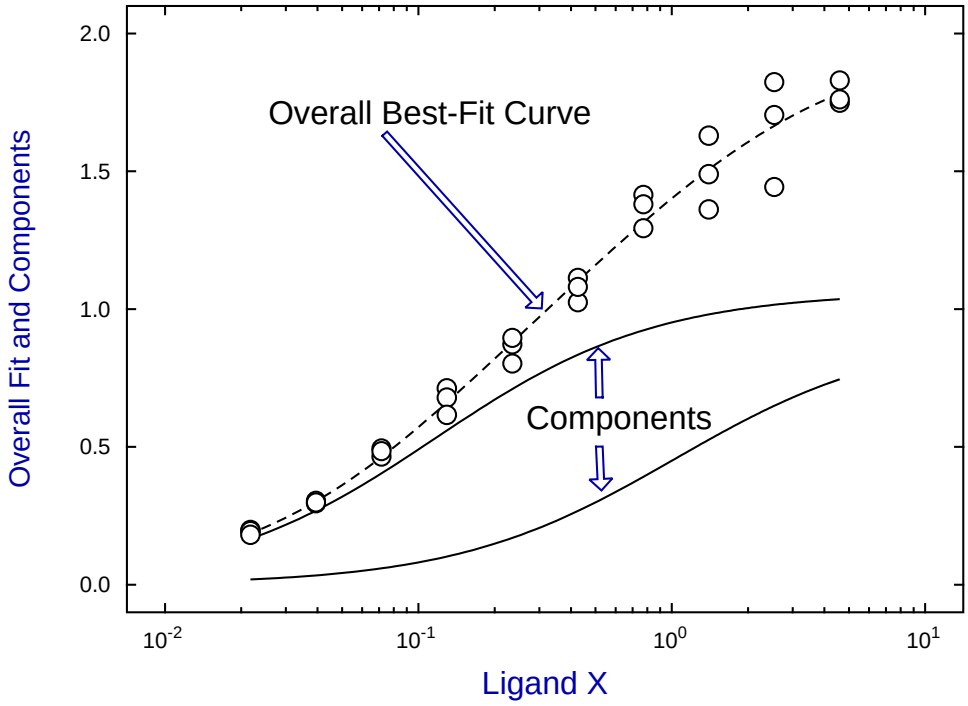
$WSSQ$ previous	86.634
$WSSQ$ current	29.952
Number of parameters previous	2
Number of parameters current	4
Number of x values	30
Akaike AIC previous	35.815
Akaike AIC current	7.9520, ER = 1.1230e+06
Schwarz SC previous	38.617
Schwarz SC current	13.557
Mallows' C_p	49.203, $C_p/2$ = 24.602
Numerator degrees of freedom	2
Denominator degrees of freedom	26
F test statistic (FS)	24.602
$P(F \geq FS)$	0.0000
$P(F \leq FS)$	1.0000
5% upper tail point	3.3690
1% upper tail point	5.5263

Conclusion based on F test
Reject previous model at 1% significance level
There is strong support for the extra parameters
Tentatively accept the current best fit model

X-semilog Plot of Fits to Models 1 and 2



X-semilog Plot for Deconvolution of Model 2



Example 2: Isotope displacement mode

When there is no appreciable kinetic isotope effect, that is, the binding and response process process is the same whether the ligand is labeled or not, this allows experiments in which labeled ligand is displaced by unlabeled ligand. Since the ratios of labeled ligand to unlabeled ligand in the bound state, and free state are equal, a modified form of high-low affinity sites equations can be used to model the binding processes. For instance, suppose that total ligand, L say, consists of labeled ligand held constant, $[Hot]$ say, and unlabeled ligand varied, $[Cold]$ say. Then the response of labeled substrate for $n \geq 1$ active sites will be given by

$$f([Cold]) = \frac{A_1 K_{a_1}[Hot]}{1 + K_{a_1}([Hot] + [Cold])} + \frac{A_2 K_{a_2}[Hot]}{1 + K_{a_2}([Hot] + [Cold])} + \cdots + \frac{A_n K_{a_n}[Hot]}{1 + K_{a_n}([Hot] + [Cold])}.$$

So, if $[Hot]$ is kept fixed and $[Cold]$ is regarded as the independent variable, then program **hlfit** can be used to fit the resulting data. In other words, cold substrate is being used as a competitive inhibitor of the saturation by hot ligand in such experiments. Note that the parameters estimated will be clear when writing the saturation with $[Hot] = u$ and $[Cold] = v$ as follows

$$\begin{aligned} f(u) &= \frac{AK_a u}{1 + K_a u} \\ g(u, v) &= \frac{AK_a u}{1 + K_a(u + v)} \\ &= \frac{\alpha\beta}{1 + \beta v} \\ \alpha &= Au \\ \beta &= \frac{K_a}{1 + K_a u}. \end{aligned}$$

This is how the estimated parameters displayed by program **hlfit** as in Table 5 must be interpreted, that is, $\hat{A}$ estimated is really an estimate for Au and $\hat{K_a}$ estimated is really an estimate for $K_a/(1 + K_a u)$.

Using the isotope displacement option in program **hlfit** with the default test file `hotcold.tf1` establishes that two sites is a statistically significant improvement over one site, and leads to the following deconvolution plot to display the best-fit curve together with the separate components.

Table 5: For best-fit order 2 isotope displacement function

Number	Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p
1	B_1	10.485	2.4284	5.5380	15.431	0.0001
2	B_2	239.06	46.121	145.11	333.00	0.0000
3	K_1	1.0124	0.19498	0.61521	1.4095	0.0000
4	K_2	0.021593	0.0063982	0.0085604	0.034626	0.0019

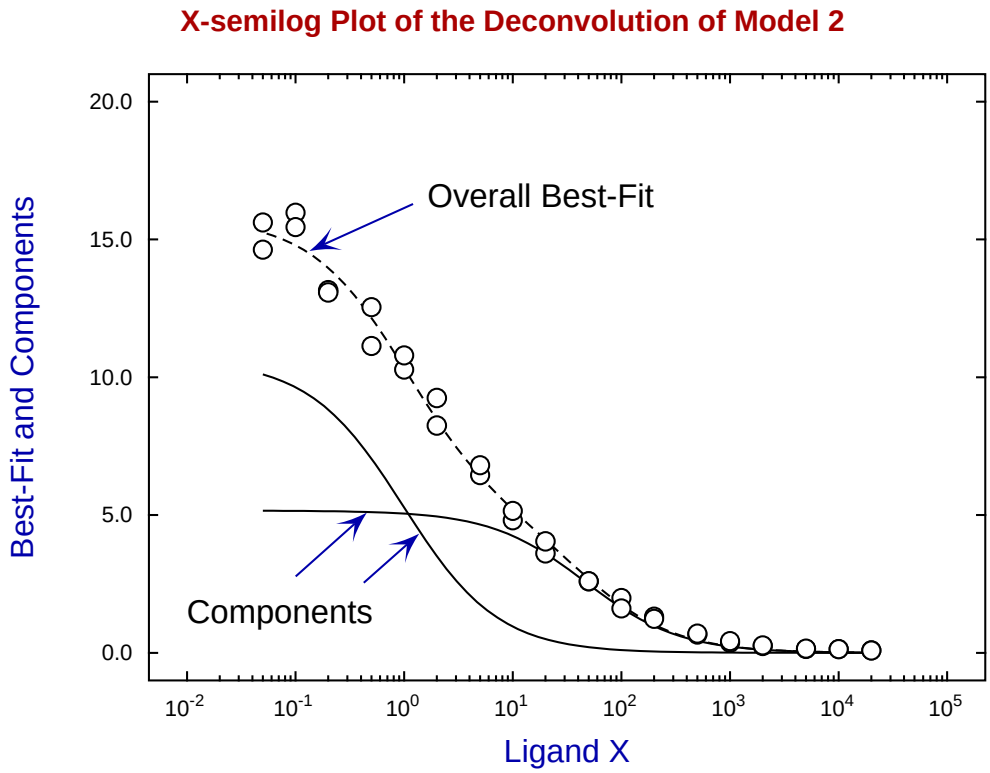
Apparent Y_{max} (i.e. $B_1K_1 + B_2K_2 + \ldots + B_nK_n$) = 15.776
Apparent K_a (i.e. x_0 where $f(x_0) - C = Y_{max}/2$) = 2.5163

Parameter correlation matrix

1			
0.5405	1		
-0.9799	-0.4559	1	
-0.7615	-0.9450	0.6712	1

Note that an important difference between using **hlfit** in this mode rather than in straightforward binding mode is that the binding constants are modified in the following sense, as previously described.

Where the actual concentration of $[Hot]$ is known it is possible to fit such data in a more satisfactory and discerning manner by using SimFIT program **qnfit**, where the $[Hot]$ can be input as a fixed constant term so that the actual amplitudes A_i and binding constants K_{a_i} can be estimated, rather than the apparent ones mentioned above.



Theory

SimFIT program **hlfit** assumes that a response is measured that depends on the fractional saturation of binding sites with possibly differing affinity. The amplitude factors A_i can be interpreted as being proportional to the population of the receptor types, possibly complicated by the situation where the fractional receptor occupancy does not give the same response for the different receptor types. Program **hlfit** also allows for the situation where there is background noise at level C that has to be estimated then substracted from the data so that the response is zero at zero ligand concentration.

The first thing to point out is that this model does not have a standard binding polynomial to act as a partition function, as it is a weighted sum of individual independent sites and can therefore only show negative cooperativity. To understand the meaning of the parameters being estimated by program **hlfit** consider the binding of a single ligand X to a protein P at equlibrium so that this is the binding process

$$P + X = PX$$

with the association constant K_a defined as

$$K_a = \frac{[PX]}{[P][X]}$$

and the fractional saturation of the protein with ligand X is $0 \le y \le 1$ defined as

$$y = \frac{K_a[X]}{1 + K_a X}.$$

However the response measured will be the fractional saturation multiplied by an arbitrary amplitude factor A, unless fractional saturation is measured when the individual amplitude factors would be nonnegative and

would have sum one. Some versions of programs **hlfit** and **qnfit** provide this feature as an additional option. In addition, in some experiments there is an unavoidable background level C which can be estimated during the fitting, or better estimated independently and then subtracted from the measured response, so that $Y(0) = 0$.

In bygone days before the advent of computers, experimentalists had to fit binding equations by plotting in transformed spaces, such as the Scatchard plot, and then extrapolating to estimate slopes and intercepts, but thankfully this era has long since gone. However, this does not mean that fitting such an equation by constrained weighted least squares is a simple process. It is not. In fact the case with $k = 1$ is trivial, the case with $k = 2$ is reasonable, but the cases $k > 2$ require data that is very extensive and accurate, and where the parameters are sufficiently distinct to allow model discrimination. For this particular model that requires amplitudes A_i values to be similar, but binding constants K_i to be distinct.

Program **hlfit** performs the following steps.

1. The y_i values are first weighted using $w_i = 1/se_i^2$, or used unweighted if all $se_i = 1$.
2. Using the ranges of x_i and y_i the data are transformed into internal coordinates of order unity.
3. Possible starting estimates are calculated for the parameters based on the internal coordinates, and then these are altered by adding pseudo-random perturbations until an approximate minimum value for the weighted sum of squares is located.
4. The parameters are then transformed into internal coordinates that will hopefully be of order unity to stabilize the optimization.
5. From these random starting estimates the lowest and highest possible limits are calculated, then constrained optimization is performed by the quasi-Newton technique.
6. The internal parameters are transformed back into user-space, and the Hessian is estimated at the solution point then inverted to calculate the parameter covariance matrix.
7. The order of parameters is permuted so that the subscripts for $i = 1, 2, \ldots, k$ refer to best-fit parameters in the order $A_1 \leq A_2 \leq \ldots \leq A_n$. This is to allows retrospective comparison of fits to alternative data sets.
8. The apparent (overall) A is calculated as the sum of the A_i (or $A_i K_{a_i} i$ for isotope displacement) and the apparent (overall) K_a is calculated numerically.
9. Analysis of the residuals is performed together with numerous statistical procedures to ascertain goodness of fit, parameter reliability, and model discrimination.
10. Results tables and graphs are then provided.

Program **hlfit** allows users to control the random search for starting estimates and the technique to be used for calculating the gradient vector, and should the cases with $k > 2$ be required, users can perform extensive random searches to obtain starting estimates that can be input retrospectively for manual starts. If these steps do not succeed it is time to try the SimFIT advanced curve-fitting program **qnfit**.

Although **hlfit** can be used to fit more than two classes of sites it must be stressed that this requires extremely accurate data over a large range of ligand concentration, and the automatic estimation of suitable starting estimates may have to be replaced by user-supplied estimates. In any case it will be extremely difficult to interpret binding data in terms of more than two classes of binding sites by curve-fitting alone unless there is additional experimental evidence.