Biophysical Journal, Volume 99

Supporting Material

Time-resolved NMR: Extracting the topology of complex enzyme networks

Yingnan Jiang, Tyler McKinnon, Janani Varatharajan, John Glushka, James H Prestegard, Andrew T Sornborger, Heinz-Bernd SchÃ¼ttler, and Maor Bar-peled

Time-resolved NMR: Extracting the topology of complex enzyme networks

Yingnan Jiang[1], Tyler McKinnon[1], Janani Varatharajan[1], John Glushka[1], James H. Prestegard[1], Andrew T. Sornborger[2], Heinz-Bernd Schüttler[3], and Maor Bar-Peled[1,4]

[1]Complex Carbohydrate Research Center (CCRC), University of Georgia, Athens
[2]Department of Mathematics/Faculty of Engineering, University of Georgia, Athens
[3]Department of Physics and Astronomy, University of Georgia, Athens, GA 30602
[4]Address correspondences to: Maor Bar-Peled, CCRC 315 Riverbend Rd., Athens, GA 30602 Fax: 706- 542-4412; Email: peled@ccrc.uga.edu

Supplementary data

Experimental procedures:

cDNA cloning of Arabidopsis ACS and E. coli expression and purification –Total RNA was extracted from the leaves of six-week-old *Arabidopsis* plants and used as a template to reverse-transcribe cDNA with oligo-dT as primer (*1*). The coding sequence of

Arabidopsis ACS (At5g36880) was amplified by PCR using 1 unit of high-fidelity proof-reading Platinum DNA polymerase (Invitrogen), and 0.2 μM of each forward and reverse primers: 5'- atgaaaataggatctccttcttccccg -3' and 5'- gccacatcggcaagtgcaataag -3'. The RT-PCR product was cloned into pGEM vector (Promega) and subsequently DNA sequencing was used to confirm gene identity. An Ecor I- NotI fragment (~2000 bp) containing the partial *ACS* gene without the putative signal peptide N-terminal region (aa97-743) was sub-cloned into an *E. coli* expression vector derived from pET28a.

E. coli cells harboring a pET28:*ACS* plasmid or a vector alone were cultured for 16 h at 37 ºC in LB medium (15 mL) supplemented with kanamycin (50 μg/mL) and chloramphenicol (34 μg/mL). A portion (5 mL) of the cultured cells was transferred into fresh LB liquid medium (250 ml) supplemented with the same antibiotics, and the cells then grown at 37^{o}C at 250 rpm until the cell density reached OD_{600} = 0.6. The cultures were then transferred to 25^{o}C and gene expression induced by the addition of isopropyl β-D-thiogalactoside to a final concentration of 0.5 mM. After 3 h growth while shaking (250 rpm), the cells were harvested by centrifugation (6,000 x g for 10 min at 4 ºC), resuspended in lysis buffer (10 ml 50 mM Tris-HCl pH 7.6, 10% (v/v) glycerol, 50 mM NaCl, 0.1 mM EDTA, 1 mM $MgCl_2$, supplemented with 1 mM DTT and 0.5 mM phenylmethylsulfonyl fluoride) and lysed in an ice bath by 24 sonication cycles each (10-sec pulse; 20-sec rest) using a Misonix S-4000 (Misonix incorporated, Farmingdale, New York) equipped with microtip probe. The lysed cells were centrifuged at 4 °C for 30 min at 20,000 x g, and the supernatant (termed s20) was recovered and kept at -20^{o}C.

ACS proteins were purified either on a gel-filtration column (Superdex75, 1 x 90 cm) followed by SourceQ-15 (0.5 x 5 cm, GE) (*1*), ATP-affinity column (2 ml, a kind gift from Dr. Timothy Haystead, Duke University, Durham) or various Dye columns (1 ml, Sigma).The ATP- and Dye-columns were equilibrated with 50 mM sodium-phosphate, pH 7.5. The bound proteins were eluted with the same buffer containing increasing concentrations of salt up to 0.5 M, and ACS selectively eluted with buffer containing 10 μM ATP. The fractions containing ACS activities were snap frozen in liquid nitrogen and stored in aliquots at -80ºC. The concentration of proteins was determined using the Bradford reagent kit (Bio-Rad) using bovine serum albumin (BSA) as standard.

The molecular weight of the recombinant ACS was estimated by size-exclusion chromatography using a Waters 626 LC HPLC system equipped with a photo diode array detector (PDA 996) and a Waters Millennium32 workstation. ACS (0.5 ml) or a mixture of standard proteins [10 mg each of alcohol dehydrogenase (150 kDa), ovalbumin (48.9 kDa), ribonuclease A (15.6 kDa), and cytochrome C (12.4 kDa)] were separately chromatographed at 1 ml/min on a Superdex75 column (10 mm id X 300 mm long, GE) equilibrated with 0.1M sodium phosphate, pH 7.6, containing 0.1 M NaCl. The eluant was monitored at A280 nm and fractions collected every 20 sec. Fractions containing enzyme activity were pooled and kept at -80ºC. To confirm the amino acid sequence of recombinant ACS, the purified recombinant protein was fragmented with trypsin and the resulting peptides were confirmed by MALDI TOF-MS analyses.

ACS HPLC-based assay- reactions (50 µl final volume) were performed in 50 mM HEPES-NaOH, pH 7.6, (or other buffer when indicated) containing 5 mM $MgCl_2$, 1 mM ATP, 0.2 mM CoA, and 1 mM acetate and 1.5 µg recombinant ACS. Reactions were kept at 37ºC for up to 10 min, and then terminated with an equal volume of chloroform. After vortexing (30 seconds) and centrifugation (12,000 rpm for 5 min, at room temperature), the upper aqueous phase was collected and chromatographed at 1 ml min^{-1} on a C18 reverse-phase column (4.6 mm id x 250 mm long, 5 µm, Agillent Prep-C18, or 4.6 mm id X 100 mm long, 5 µm TSKgel ODS-100V, Tosho) using an Agilent Series 1100 HPLC system equipped with an autosampler, diode-array detector and ChemStation software. Chromatography conditions were 35 min linear gradient from 98% A/2% B to 100% B [where A is (20 mM t-butylamine-H_3PO_4 pH 6.6) and B is (20 mM t-butylamine -H_3PO_4 pH 6.6 with 20% acetonitrile)] followed by 10 min at 100% B, 3 min to 50% A/B and 10 min pre-equilibration with 98% A/2% B. Nucleotides were detected by their UV absorbance using a Waters or Agilent photodiode array detector. The maximum absorbance for adenosine-nucleotides was 253 nm. The peak areas of analyses were compared to calibration curves of an internal standard. Stds (1 mM each ATP, AMP, acetate, CoA, AcCoA) were prepared in the same buffer, temperature and pH as enzymatic reactions.

Kinetics-To determine the kinetic parameters for ATP, acetate, and CoA, the experiments were conducted by varying the concentration of one substrate while the other was saturated. The catalytic activity of ACS was assayed at 37ºC for 5 min using HEPES-NaOH pH 7.6, containing $MgCl_2$ (5 mM), variable concentrations of ATP (40 µM to 2 mM), a fixed concentration of acetate (2 mM) and CoA (2 mM), and 0.2 µg recombinant ACS (3 pmol). In a separate series of experiments reactions were performed with a fixed amount of ATP (2 mM) and variable concentrations of either CoA or acetate (20 to 400 µM). Enzyme velocity data of the amount (µM) of AcCoA produced per second per µg enzyme, as a function of substrate concentrations was plotted. The Solver tool (Excel version 11.5 program) was used to generate best-fit curves calculated by nonlinear regression analyses, and for calculation of Vmax and apparent Km.

Enzyme properties of ACS-To characterize the properties of recombinant ACS, the activity was tested under a variety of conditions: with various buffers, at different temperatures, different ions, or with different potential inhibitors. For the optimal pH experiments, 1.5 µg recombinant enzyme was first mixed with 5 mM $MgCl_2$, 1 mM ATP, 1 mM CoA and 50 mM of each individual buffer (Tris-HCl, phosphate, MES, MOPS, or HEPES). Assays were then initiated after the addition of 1 mM acetate. Inhibitor assays were performed under standard assay conditions except for the addition of various additives (DTT, nucleotides) to the reaction buffer. These ACS assays were incubated for 30 min at 37ºC, and were subsequently terminated by heat (1 min at 100ºC). After cooling and chloroform extraction, the reaction was separated by chromatography and the amount of AcCoA formed was calculated from HPLC UV spectra. For the experiments aimed at defining the optimal temperature the ACS assays were performed under standard assay conditions except that reaction were incubated at different temperatures for 10 min. Subsequently, the activity was terminated (100ºC) and the relative activity was measured after chromatography. For the experiments aimed at determining if ACS

required metals, assays were performed with buffer, ATP, CoA, acetate and with a variety of ions. After 10 min at 37ºC incubation the assay was terminated by heat. The amount of AcCoA formed was calculated from HPLC UV spectra.

Deconvolution of superimposed peaks-In the "multiple peak spectral fitting' (MPSF) approach described below, metabolic time-series spectra are to be analyzed as linear superpositions of the standard spectra of individual metabolites. However, between the measurement of the standard spectra and the measurement of the metabolic time-series spectra NMR resonances tend to shift due to small changes in the physico-chemical environment. As a pre-processing step, we therefore have to align the peak positions of each observed standard to match the corresponding peak positions in the time-series spectra. The resulting "aligned spectral standards" will then be used to represent the respective metabolite's contribution to the time-series superposition spectra.

In the simplest version of our alignment procedure, we utilize only certain well-resolved, isolated peaks in the time-series spectra, to be referred as "diagnostic peaks", which must be chosen for each metabolite so as to *not* overlap with any peaks of any other metabolites. The alignment is then applied separately to each diagnostic peak of each observed metabolite. This is implemented by representing the "target" diagnostic peak in the time-series spectra as a superposition of a large set of "shifted diagnostic peaks", generated by applying successive frequency shifts, in small increments, to the original diagnostic peak extracted from the observed standard spectrum. That is, formally $\mathbf{e} = \mathbf{L}\mathbf{a}$ where the experimental data vector $\mathbf{e}$ comprises the target diagnostic peak signal of the time-series spectra; each column of the "library" matrix $\mathbf{L}$ contains a shifted diagnostic peak generated from the standard spectrum; and $\mathbf{a}$ is the vector of superposition amplitudes. The diagnostic peak shifts that make up $\mathbf{L}$ are generated in a window (-W,W) about the original peak from the standard beginning at location –W, then successively shifting the peak by one NMR frequency grid point (*i.e.*, $\sim 1.3 \times 10^{-4}$ ppm for the data presented here) up to W. The usual method for determining $\mathbf{a}$ would be to invert (or find a pseudoinverse of) the matrix $\mathbf{L}$ and solve: $\mathbf{a} = \mathbf{L}^{-1}\mathbf{e}$. However, because $\mathbf{a}$ is not otherwise constrained, this approach does not localize the peak well. Furthermore, oscillations and negative peaks are found in the estimate of $\mathbf{e}$, which is unphysical. For these reasons, we use a rank one variant of an optimization method called Non-Negative Matrix Factorization (NMF) (*2-3*). This method enforces the constraint that the vector $\mathbf{a}$ is non-negative. In other words, all of the peak amplitudes are positive. This type of factorization minimizes the distance $|\mathbf{e} - \mathbf{L}\mathbf{a}|$. For our results, we used Lee and Seung's original multiplicative update rule:

$$\mathbf{a}_i =: \mathbf{e}_i \, (\mathbf{L}^T \mathbf{e})_i / (\mathbf{L}^T \mathbf{L}\mathbf{a})_i \qquad (1)$$

where superscript T indicates the transpose. As illustrated below, the resulting amplitudes $\mathbf{a}$ exhibit a sharp maximum when plotted as a function of the incremental peak shifts applied to the standard. The corresponding maximum-amplitude shifted diagnostic peak gives the best alignment to the target diagnostic peak in the time series and it is therefore used as the "aligned spectral standard" to represent the metabolite's spectral contribution to the time-series spectra in the particular frequency window where it resides. Alignment errors are assessed by picking random initial conditions for the NMF minimization procedure and generating a probability distribution of possible alignments.

The same alignment procedure can be applied, separately and independently, to each identifiable diagnostic peak of each metabolite for which a standard spectrum has been recorded. This allows us to incorporate multiple shifted diagnostic peaks, each from a different spectral frequency window, into a more detailed and more informative aligned spectral standard. One should expect, and we have explicitly confirmed, that inclusion of a larger number of peaks in the aligned spectral standard significantly reduces the noise in the resulting MPSF-generated concentration estimates. Errors for this multiple window procedure are generated in the same way as for individual windows.

The foregoing NMF-based alignment procedure can be further generalized to deal with metabolites which do not exhibit any isolated diagnostic peaks. In this case, the NMF is employed to simultaneously fit multiple standards to the time-series data in a given frequency window. This generalized multiple-standard alignment approach can then also be applied to any other spectral frequency windows of interest where multiple metabolites may exhibit overlapping standard spectral weight. We have found that this approach generally gives low noise results, specifically by way of reducing noise in the resulting MPSF-based concentration estimates. All MPSF concentration data reported here were therefore generated via multiple-standard alignment in each given frequency window.

The Ensemble Network Simulation (ENS):

Input Data Sets and Models-From the proton NMR spectral time-series of both the forward reaction (FR) and backward reaction (BR) experiments, time-dependent concentrations of small-molecule metabolite compounds acetate, ATP, CoA, acetyl-AMP (also referred to as compound "Q" for short), AcCoA, and AMP were extracted for up to 300 observation times, spread out approximately equidistantly over an ~73 minute reaction and NMR observation time interval, by two different methods: the MPSF and peak integration with background subtraction (PIBS) approach, as described in the *Experimental Procedures* section. In addition, the peak integral time-series of an NMR resonance, representing the concentration sum of three adenosine-containing compounds (ATP, CoA and AcCoA), was included in the PIBS data set. Each of these two time-series data sets (MPSF and PIBS) was then used as input into a series of ensemble network simulations (ENS) following Battogtokh *et al.* 2002 and Yu et al. 2007 to discriminate between four different hypothesized reaction network topologies of the ACS enzymatic pathway [Figure 6(A)] (*4-5*).

Ensemble Probability Distribution-In each ENS, both FR and BR time series data for all ~300 observation times and all observable compounds (acetate, CoA, ATP, Q, AcCoA, and AMP), as well as the observed concentration sum ATP+CoA+AcCoA for the PIBS data set, were incorporated into an ensemble probability distribution function $Q(\theta,\phi) = \Omega^{-1} \exp[-\chi^2(\theta,\phi)/2]$ with normalization factor Ω and a χ^2-function,

$$\chi^2(\theta,\phi) = \sum_{j=1\ldots J} [Y_j - \phi_{c(j)} - F_j(\theta)]^2 / \sigma_j^2 \,. \qquad (2)$$

Here, $\theta \equiv (\theta_1, \ldots \theta_M)$ denotes the vector of all unknown rate coefficient and unknown initial concentration parameter variables in the model, with M=15, 21, 21, and 23 in Models 1, 2, 3, and 4, respectively. Of these, M_s=8 are initial concentration variables and M_r=7, 13, 13, and 15 are rate coefficient variables for models 1, 2, 3, and 4, respectively. The index j=1,…J labels all data points from all experiments, all observable compounds

and all observation times, with J=3878 and 2995 in the PIBS-processed and MPSF-processed data set respectively. Y_j is the natural log of the observed concentration, measured in NMR peak area units; σ_j is the corresponding experimental standard deviation (SD) of Y_j; $F_j(\theta)$ is the corresponding predicted natural log molar concentration, obtained by solving the model's kinetic rate equation system for given model parameter vector θ; and $\phi_{c(j)}$ is the natural log of the unit conversion factor required to convert the model's molar concentration units into the experimental NMR peak area units in which the concentration for data point j is measured.

Unit Conversion Factors and Scale Factor Classes-The unit conversion factors required to convert the model's molar concentration units into experimentally observed NMR peak area units may differ from one experiment to another, due to changes in experimental detection conditions; and they may also differ between resonances within the same NMR spectrum, due to differential T_1 relaxation times. To determine (or at least constrain) these conversion factors in the ENS, the set of all experimental data points j is partitioned into "scale factor classes" (SFCs) (*5*), such that all j sharing a common unit conversion factor (from molar to NMR resonance peak area) are assigned to the same SFC, with different SFCs labeled by an index c=1,… C (*5*) The $\phi \equiv (\phi_1, \ldots \phi_C)$ then denotes the vector of the natural log conversion factors, ϕ_c, and c(j) in Eq.(2) denotes the SFC which comprises data point j. The ENS treats each ϕ_c as an independently adjustable model parameter variable, on an equal footing with the θ-variables. For the MPSF and the PIBS data set, two or, respectively, three SFCs with independently adjustable ϕ_c (*i.e.*, C=2 and C=3) were introduced. For the MPSF data set, the SFC c=1 comprises all data points j for all compounds observed in the FR experiment; and c=2 comprises all data points j for all compounds in the BR experiment. For the PIBS data set, the SFC c=1 comprises all data points j for all compounds observed in the FR experiment, as well as compound Q in the BR experiment; c=2 comprises all data points j for all compounds except AcCoA in the BR experiment; and c=3 comprises data points j for compound AcCoA in the BR experiment.

The ϕ-variables are constrained in the ENS by fixing the model initial (t=0) concentrations to the value 1.0mM (which is independently known experimentally, from the titration of the input reactants) for Acetate in the FR, and for AMP and PPi in the BR experiment. For other input reactant compounds, the initial concentrations may not have been well quantitated experimentally by titration, in either FR or BR experiments, and they were therefore treated as MC θ-variables in all ENSs. For observable output product compounds (AcCoA and AMP in FR; and acetate, CoA and ATP in BR) the initial concentrations were also treated as θ-variables, each with an upper bound of 0.25 mM, to provide for ENS correction of small, systematic background subtraction errors in the MPSF or PIBS data sets. ACS enzyme initial concentrations were fixed at the titrated values of 0.0008 mM for both the FR and the BR experiment. The initial concentration values of the hypothesized enzyme-complex compounds occurring in the various models were all fixed to zero for both FR and BR.

Averaging over the ϕ-variables can be carried out analytically, since the ϕ-variables are normally distributed in $Q(\theta,\phi)$ for fixed θ. One thereby obtains an "effective", ϕ-averaged probability distribution, with an "effective" χ^2-function, denoted

respectively by $Q(\theta)$ and $\chi^2(\theta)$ below, where $Q(\theta) = \text{const} \times \exp[-\chi^2(\theta)/2]$. Due to the normal distribution of the ϕ-variables in $Q(\theta,\phi)$, this is mathematically equivalent to minimizing $\chi^2(\theta,\phi)$ with respect to ϕ at fixed θ. It is also equivalent to replacing the independent ϕ-variables by their "conditional" means, *i.e.*, by ϕ averaged over $Q(\phi|\theta) \equiv Q(\theta,\phi)/Q(\theta)$, given a fixed θ. This conditionally averaged ϕ-vector at fixed θ is denoted by $\phi(\theta)$ for short. The effective $\chi^2(\theta)$ thus equals the full $\chi^2(\theta,\phi)$-function, minimized with respect to the ϕ-variables. The Monte Carlo (MC) guided random walk procedure discussed below is then implemented for this effective $Q(\theta)$, *i.e.*, it randomly varies only the θ-variables, guided by the effective $Q(\theta)$, with ϕ-variables replaced by their conditional means $\phi(\theta)$. All χ^2-values and means shown and discussed here and in the main text of this article are (based on) effective χ^2-values. In all figures comparing experimental concentration time series data to ENS-based model predictions [Figures 6(B) and S2(A)], the experimental data $\exp(Y_j)$ have been re-scaled into model units by multiplication with $\exp(-\langle\phi_{c(j)}\rangle)$. That is, the experimental data points shown, in units of mM, represent $\exp(Y_j - \langle\phi_{c(j)}\rangle)$, with $\langle\phi_{c(j)}\rangle$ denoting the mean of $\phi_{c(j)}$ over $Q(\theta,\phi)$ or, equivalently, the mean of $\phi_{c(j)}(\theta)$ over $Q(\theta)$.

Integrating out the ϕ-variables by analytical means does of course not eliminate their ensemble fluctuations prescribed by the underlying probability distribution $Q(\theta,\phi)$. Even though the ϕ-variables are not independently varied by the MC random walk, the ensemble ϕ-fluctuations, along with the θ-fluctuations, do still contribute to the ensemble uncertainties of predicted observables, such as the uncertainty bands of the concentration time series shown in Figures 6(B) and S2(A). Otherwise it might appear paradoxical, for example, that model 3 could have the lowest converged χ^2-values, as seen in Figure 6(C), and at the same time have the largest width of concentration uncertainty bands, as seen in Figures 6(B) and S2(A). In the absence of the ϕ-variables, the model with the widest concentration uncertainty bands would also be expected to have the largest (not smallest!) χ^2-values, since the uncertainty band widths directly correlate with the rms magnitude of the "zero-ϕ residuals" $R_j := (Y_j - F_j(\theta))$ and the latter would also be the residuals contributing to χ^2. However, in the presence of the ϕ-variables the rms values of the "χ^2-residuals" in Eq.(2), $r_j := (Y_j - \phi_{c(j)} - F_j(\theta))$, are not necessarily correlated in an obvious way with the concentration uncertainty band widths. The uncertainty band widths are again directly correlated with $R_j = (Y_j - F_j(\theta))$, but now $R_j = r_j + \phi_{c(j)}$ and $\langle R_j^2\rangle = \langle r_j^2\rangle + \langle\phi_{c(j)}^2\rangle + 2\langle r_j\,\phi_{c(j)}\rangle$. So, when expressed in terms of the mean squares of χ^2-residuals, the corresponding mean square of the zero-ϕ residuals is modified both by the square of the relevant ϕ-variable and the cross-correlation between the $\phi_{c(j)}$-variable and r_j. The ϕ- and θ-variables can then be coupled by $Q(\theta,\phi)$ in such a manner that, on average, the mean χ^2 is decreased by the coupling, while the concentration uncertainty band widths are increased, due to the additional, positive $\langle\phi_{c(j)}^2\rangle + 2\langle r_j\,\phi_{c(j)}\rangle$-contribution. This is in fact the case in the models studied here, and most strongly so in model 3.

Experimental Standard Deviations- Experimental SD values σ_j are required as inputs to Eq. (2). They were estimated separately for each compound in each experiment with two independent methods: either by (i) least-squares fitting a linear function $y=at+b$ to the experimental concentration data y_j (in NMR peak area units) over a short (<5min)

time interval and calculating their rms deviation $\sigma^{(rms)}$ from the linear fit; or by (ii) generating a distribution of concentrations from the alignment distributions found in the MPSF procedure and calculating standard deviations of the set of resulting concentrations. Both of these procedures generated similar error bars and resulted in $\sigma^{(rms)}$-values which are roughly independent of time over the ~73min NMR observation time interval for each compound and experiment. We then used $\sigma_j=\ln(1+3.0\times\sigma^{(rms)}/y_j)$ for the SD of the log concentration $Y_j\equiv\ln(y_j)$. Multiplying $\sigma^{(rms)}$ by the additional factor of 3.0 here ensures that we are obtaining a very conservative upper limit for σ_j. Using larger σ_j speeds up MC equilibration and it also helps to avoid overfitting of the data. Consequently, we are likely overestimating the uncertainties in any ensemble predictions based the resulting distribution $Q(\theta,\phi)$, since increasing σ_j tends to "widen" the distribution.

Constraints by Sub-Noise Metabolites-The NMR spectra did not show any evidence of any detectable (*i.e.*, above-noise) amounts of the intermediate metabolite acetyl-AMP (Q) being released as a free non-enzyme-bound species during either the FR or BR experiment. While this observation does of course not rule out the production of small sub-noise levels of free acetyl-AMP, it *does* impose an upper limit on the free acetyl-AMP levels; and this upper limit in turn could impose a potentially important constraint on any hypothetical pathway model. To incorporate this constraint into the ENS, we included in $\chi^2(\theta,\phi)$ for the PIBS simulations a series of acetyl-AMP "synthetic" experimental data $Y_j\equiv\ln(y_j)$, at 300 time points each for both the FR and the BR experiment, with all y_j set to the mean $\sigma^{(rms)}$ from compounds detected in the FR experiment and with a large SD for Y_j, of $\sigma_j=2.0$. This has the effect of allowing model parameterizations θ with possible Acetyl-AMP production at or below the noise-limited NMR detection levels, while constraining the ensemble distribution so as to suppress models with Acetyl-AMP levels rising significantly above NMR noise.

For the MPSF data set, separate simulations were performed for each model both *with* and *without* such synthetic "Q-constraint" data included in $\chi^2(\theta,\phi)$. We found that the MPSF results were generally not affected by inclusion or omission of these synthetic data. Specifically, the rank-ordering of the models by their respective MC-converged χ^2–values does not depend on the presence or absence of the synthetic data. Figure 6 and Figure S2 only show the MPSF and PIBS results from the simulations performed *without* synthetic Q-constraint data.

Monte Carlo (MC) Protocol-During the simulation, all unknown model parameter variables θ_m are confined to intervals $[\theta_m^{(lo)}, \theta_m^{(hi)}]$ with very conservative low- and high-limits $\theta_m^{(lo)}$ and $\theta_m^{(hi)}$, respectively. At the start of the simulation, each θ_m is randomly initialized with $\ln(\theta_m)$ drawn from a uniform distribution on $[\ln(\theta_m^{(lo)}), \ln(\theta_m^{(hi)})]$. Starting from this random θ-initial, a MC equilibration of 5000 MC sweeps is performed. As described in (*5*), each MC sweep consists, with equal probability, of either M single-θ_m or M global-θ Metropolis updating steps, controlled by the ensemble probability distribution $Q(\theta)$. MC equilibration is then followed by MC sample accumulation, consisting of another 1000 or 5000 MC sweeps for the MPSF- or PIBS-processed experimental data, respectively, with the random θ at the end of each sweep being

collected into the MC sample. Ensemble averages, SDs and parameter distributions are estimated by the corresponding averages, SDs and distributions over the MC sample, including, for example, MC averages of the kinetic rate equation solutions for the time-dependent metabolite concentrations. This MC process of random θ-initialization, 5000 equilibration and 1000 (for MPSF) or 5000 (for PIBS) accumulation sweeps is repeated 20 or 10 times, for the MPSF- or PIBS-processed data sets, respectively, each MC (re-)start with a different random θ-initialization. Of these 20 or 10 MC re-starts, only those converged with an accumulated MC average $\chi^2<\chi_{max}^2=6000$ are collected into a "grand sample", for each of the four models, simulated with the MPSF input data set. The choice of this χ^2-cut-off, χ_{max}^2, is explained further below.

For the PIBS input data set the overall quality of fit was significantly worse for all four models, with the best converged χ^2-values from all MC re-starts being about 4 times larger than for the MIPS input data set, as illustrated below in Figure S2(B). We therefore had to increase the collection tolerance for the PIBS simulations, to $\chi_{max}^2=20000$, in order to be able to collect any MC re-starts at all into the grand sample. As discussed below, the PIBS time series data cannot be adequately fitted by any of the four network models considered here. However, as discussed below and in the main text of this article, this is a deficiency of the PIBS data analysis procedure, and not *per se* a deficiency of the four models. We are showing the PIBS-based ENS results here only to further illustrate the inadequacies of the PIBS procedure and to contrast it with the MPSF approach. The latter extracts sufficiently reliable concentration time series data to allow for a meaningful ENS-based analysis and network model discrimination; the former does not.

The grand sample is used to estimate the MC statistical sampling error and to obtain the ensemble grand sample averages and SDs, as shown in Figure 6(C) and Figure S2(B). For each of the four models, simulated with each (MPSF and PIBS) input data set, the MC re-start with the lowest final χ^2-value was then also used for the MC trajectories in Figure 6(C) and Figure S2(B), showing $\chi^2(\theta)$ *vs.* MC sweep number.

The number of experimental data points J included in χ^2, Eq. (2), sets the χ^2-scale for an acceptable fit. Assuming approximately normally distributed experimental data, we expect an acceptable fit to have a χ^2-value of about $\chi_{exp}^2=J$, *i.e.,* $\chi_{exp}^2 \sim 3000$ for the MPSF and $\chi_{exp}^2 \sim 3900$ for the PIBS data sets, respectively. If the actual χ^2–values of the best model fit are substantially larger than $\chi_{exp}^2=J$ we should reject the underlying network model as incompatible with the data.

Likewise, if the actual ensemble mean χ^2-value of a MC re-start substantially exceeds $\chi_{exp}^2=J$ then we should reject such a re-start from the MC grand sample. The number of experimental data points J thus also sets the scale of our χ^2-cut-off, χ_{max}^2, for acceptance or rejection of the specific parameter (θ-) space region explored by an MC re-start. We have chosen as $\chi_{max}^2 = 2\chi_{exp}^2 = 2J \sim 6000$ for the MPSF data. The expected ensemble standard deviation of χ^2 is also controlled by the number of experimental data points J, with $\sigma[\chi_{exp}^2] \sim J^{1/2} \sim 55$. The imposed cut-off χ_{max}^2 exceeds the expected mean χ^2-value, χ_{exp}^2, by about (6000-3000)/55 ~ 55 standard deviations. So only MC re-starts with converged mean χ^2-values within less than 55 expected standard deviations from the expected mean, χ_{exp}^2 have been included into the MC grand sample for the MPSF data set.

Some MC re-starts can in fact produce converged mean χ^2-values significantly exceeding those of the "better" (*i.e.*, lower-χ^2) re-starts, and thus exceeding $\chi_{\mathbf{exp}}^2$. This occurs when the MC random walk through θ-space gets trapped in a high-χ^2 local minimum of the $\chi^2(\theta)$-landscape. By performing repeated MC re-starts, each with a completely new random θ-initialization, instead of just one very long single MC run, we are reducing the probability of accidentally being trapped in only one single "bad" high-χ^2 local minimum while missing "good" θ-regions with substantially lower χ^2 –values. As in any non-linear data fitting procedure, we can of course never completely rule out the possibility of missing the "best", *i.e.*, absolute $\chi^2(\theta)$-minimum. However, the MC re-starts actually collected into our grand sample have mean converged χ^2-values that are typically within less than ~1-2 times the expected standard deviation $\sigma[\chi_{\mathbf{exp}}^2]$ from each other, suggesting that the ENS is indeed exploring θ-regions with close to absolute-minimal $\chi^2(\theta)$.

Parameterization and Prediction Uncertainties-R2Q4.2>> If mean converged χ^2-values were found to be significantly smaller than $\chi_{\mathbf{exp}}^2$ this would indicate a serious over-fitting of the data. However, the ENS approach effectively prevents this from happening. In contrast to conventional maximum likelihood methods, the ENS algorithm does not endeavor to find *the best* θ, having the minimum χ^2-value and giving the best fit to the data. Rather, for an acceptable model, the ENS generates a representative sample of *all possible* θ that are reasonably, and about equally well, consistent with the data $Y_{\mathbf{j}}$, within the constraints imposed by the overall available range of model predictions $F_{\mathbf{j}}(\theta)$. The ENS thus explores the entire "uncertainty cloud" in θ-space which comprises all θ having acceptably low, but not necessarily minimal $\chi^2(\theta)$. In the course of doing so, the ENS automatically detects the presence of large θ-uncertainties wherever they arise. The ENS also translates θ-uncertainties (and, as discussed above, ϕ-uncertainties) into corresponding uncertainties for predicted observables, as indicated, for example, by the uncertainty bands in the concentration time series shown in Figure 6(B) and S2(A).

Hence, a model which would over-fit the data in a maximum likelihood setting can still be reasonably analyzed by the ENS approach, and utilized to give meaningful, if uncertain, predictions. In an ENS analysis, such an "over-fitting" model will exhibit large θ-uncertainties, at least along some sub-manifolds of θ-space. These θ-uncertainties may, or may not, translate into large prediction uncertainties, depending on the model and on the specific observable quantity that is being predicted. In the context of kinetic rate equation network models it has been found (*4,5,10*) that large θ-uncertainties frequently *do not* always result in large prediction uncertainties for concentration time series observables. Concentration time series for some, if not all, relevant molecular species can often be predicted with reasonably tight uncertainty bands in spite of the fact that the underlying θ-uncertainties may be very large. This happens when the predicted observable is largely insensitive to θ-variations along those sub-manifolds in θ-space that support the large uncertainty clouds. Likewise, despite large θ-uncertainties, the ENS can still identify statistically significant differences in converged χ^2–values, and thereby discriminate, between different models.

This capability of the ENS approach to make meaningful model predictions and allow for significant model discrimination, *in spite of* large model parameterization

uncertainties, is also very much in evidence for the models and for the MPSF data set investigated here. As illustrated by the results for the ensemble means and ensemble standard deviations in Table S2, the MPSF data constrain the rate coefficient θ-variables in each of the four models only to within an order of magnitude at best. Yet the differences in converged-χ^2 values between these models are statistically significant, as discussed below.

Also, the ENS makes meaningful concentration time series predictions here, with reasonably tight uncertainty bands not only for all of the observed compounds, but also for at least some of the species for which there is no information at all in the NMR data. This is illustrated, for example, by the PPi time series predictions shown in Fig. S2(A). There are no experimental data for this compound; and yet, despite the large relative θ-uncertainties shown in Table S2 (±100% or larger!), the PPi uncertainty bands are no more than 20-30% of predicted terminal concentration. The PPi time series prediction is thus sufficiently precise to allow for a meaningful test by future experiments, such as ^{31}P-NMR.

If one were to attempt a maximum-likelihood analysis, such as least-squares fitting, for any of the four models considered here, using the current MPSF data set, the results would be meaningless, due to over-fitting. Even if a unique parameterization θ, with absolute-minimal $\chi^2(\theta)$, could be identified any model prediction based on such a single "best" θ should not be trusted. The "best" θ here is embedded in a large uncertainty cloud of "many" other, almost equally "good" parameterization choices. Any θ within this uncertainty cloud would give essentially the same goodness of fit to the data, as quantified by the χ^2-value. The standard deviations reported in Table S2 do in fact provide us with some rough measure of the "extent" of this cloud in θ-space. The only meaningful way to make any predictions in this situation is thus to locate and explore the entire cloud, not just a single θ within it, and to then translate the cloud from θ-space into probability distributions of predicted time series and χ^2-values. That, in essence, is what the ENS algorithm accomplishes.

Model Discrimination.-Visual inspection of residuals is not a reliable gauge of the systematic deviations between model and data in this case where one is dealing with complex, heterogeneous data sets with multiple species being simultaneously fitted by the same model. The χ^2-function, in the other hand, provides a quantitative measure of both random and systematic deviations between model and experiment. The different χ^2-values of the four models compared here reflect the differences in their systematic deviations from the experimental data. These deviations are well outside of their statistical χ^2-uncertainties.

The expected χ^2-values due to random fluctuations in the data are of order $\chi_{exp}^2 \sim J$ where J~3000 is the number of experimental (MPSF) data points included in χ^2. For the best-fitting model, model 3, simulated with the MPSF data set, the actual χ^2-values, of order 4400, are consistent with $\chi_{exp}^2 \sim J$. The expected standard deviation of χ^2 is of the order $\sigma[\chi_{exp}^2] \sim J^{1/2} \sim 55$. This value of $\sigma[\chi_{exp}^2]$ is also roughly consistent with the rms statistical fluctuations of χ^2, as seen in the Monte Carlo trajectories shown in Figure 6(C). The difference in χ^2-values of the two best-fitting models, Model 2 and Model 3, is of

order $\Delta\chi^2$~480. This $\Delta\chi^2$ exceeds, at least 8-fold, their standard deviation $\sigma[\chi_{exp}^2]$. This suggests that the χ^2-difference between the two models is indeed statistically significant.

The four models being compared here differ in their number of degrees of freedom, that is, in the number M of their parameter variables in θ that are being randomly varied during the MC simulation. Statistical information criteria (*8,9*) such as the Akaike information criterion (AIC) and the Schwarz-Bayes information criterion (SBIC) provide systematic approaches to account for these differences in model complexity when comparing and ranking such models in terms of their χ^2-values. According to either criterion, one should modify the χ^2-function by adding an M-dependent penalty term for the θ-space dimension M, resulting in "IC-functions" given by

$$AIC = \chi^2 + 2M \qquad \text{for Akaike}$$

and

$$SBIC = \chi^2 + 2M \ln(J) \qquad \text{for Schwarz-Bayes}$$

where J is again the number of data points in χ^2. Models with different M-values should then be ranked according to their IC-function values instead of their χ^2-values. However, for our four models, the differences in their M-values, and hence differences of the foregoing M-penalty terms in their AIC- and SBIC-functions, are either zero or small compared to the differences, $\Delta\chi^2$, in their χ^2-values, as illustrated in Figure 6(C). Taking into account the M-penalties, from either AIC or SBIC, will therefore not change the rankings based on χ^2-values. Specifically, the M-values of model 1, 2, 3, and 4 are M=15, 21, 21, and 23, respectively. Between the second-best and the best model, 3 and 2, we thus have no M-penalty difference at all: $2\Delta M=0$, using AIC, and $2\Delta M \ln(J)=0$ using SBIC. Between the third- and second-best models, 4 and 3, we have a positive M-penalty difference which only favors model 3 (and model 2!) over model 4 more strongly, with $2\Delta M=+4.0$, using AIC, and $2\Delta M \ln(J)=+32.0$, using SBIC. Between the fourth- and the third-best model, 1 and 4, we have a negative M-penalty difference, with $2\Delta M=-16$, using AIC, and $2\Delta M \ln(J)=-128$, using SBIC. However, either M-penalty difference, which tends to favor model 1 over model 4, is still much smaller in magnitude than the positive χ^2-difference, $\Delta\chi^2$~+500. Thus, $\Delta\chi^2$ favors model 4 over model 1 much more strongly than the M-penalty difference favors 1 over 4; and it does so with a net AIC- or SBIC-difference that is still well outside of the χ^2-uncertainties, $\sigma[\chi_{exp}^2]\sim J^{1/2}\sim 55$. <<R2Q4.2

Limitations and Extensions of NMF and ENS Approaches-As currently implemented, the ENS method requires assumptions of specific network model topologies to be tested against the NMR time series data. Thus, our present ENS analysis focuses on the comparison and χ^2-ranking of just a small candidate set of reasonable, but ultimately *ad hoc* model network topologies. For such a set of selected network topologies, the ENS then provides the useful capability to discriminate between the competing models. This scenario is in fact frequently encountered in the study of metabolic systems. If some candidate set of possible model topologies has already been identified by prior experiments, the ENS can identify the best candidate, or at least narrow down the possible choices, based on consistency with the data, as more experiments are being performed and more data are added.

However, in the current approach, we can of course not rule out that some network topology, omitted from the initial candidate set, may give better agreement with the data than the best candidate identified in the set. For example, even the "best" model in our candidate set, model 3, does still exhibit systematic deviations between ensemble mean and experiment for the early-time Acetate data shown in Figure S2(A). These systematic deviations account for the fact that the actual ENS mean χ^2-value exceeds the expected $\chi_{exp}^2 \sim J \sim 3000$ by about 1400 in model; and they, in fact, provide the basis for ranking of models. Conceivably, a more elaborate model topology may be found with a mean χ^2-value closer to (but not, see below, less than!) χ_{exp}^2, whereas the ENS approach implemented here determines the best topology only within a fixed candidate set.

It would therefore be useful to extend the ENS approach so that variations of network topology can be systematically explored by the simulation algorithm itself, without being constrained by the *ad hoc* selection of an initial limited, fixed candidate set. The limitation to fixed network topologies can indeed be overcome by a *variable-topology* ENS method where discrete (binary) random variables are included in θ to model, and randomly vary by MC, the presence or absence of putative reactive links in the network. This will be the focus of future developments of this methodology.

Extensions of the NMF and ENS approaches can also be developed to analyze metabolic systems where only incomplete subsets of participating metabolites can be identified from databases of spectral standards. NMF can then help in such cases to detect yet unidentified metabolites by subtracting out the already known spectral contributions. Furthermore, already in its present incarnation, ENS does not require complete time series data for all molecular species, nor even for all metabolites. This is illustrated already by the model simulations presented here: no time series data were available for the free enzyme or hypothesized enzyme-substrate complexes that are assumed in the models. Yet, the ENS is able to infer the concentration time series of all of these un-observed species. Typically, as one might expect, the ENS relative uncertainty bands of such unobserved species (as a percentage of concentration) are much larger than for the observed compounds. Lastly, the ENS can also be extended to extract best-guess spectral signatures of unknown metabolites from direct simulations of full spectral time series in both fixed- and variable-topology settings.

Results:

Cloning and biochemical characterization of acetyl-CoA synthetase from Arabidopsis:

Two alternative gene models exist for Arabidopsis ACS with respect to the translation start site and perhaps its possible sub cellular localization. The first model predicts the At5g36880.1 protein corresponding to aa (1-743) and its transcript is supported by numerous cDNA's and EST's sequencing projects (GenBank). This protein version is predicted to localize in chloroplast based on the algorithm developed by K. Nakai (www.psort.org) with a putative chloroplast transit peptide domain (spanning aa 1-to 85). The second model At5g36880.2, predicts a shorter protein (693 aa) lacking the first 51 aa. The latter version was predicted to reside in the secretory system or in other organelles such as peroxisome. Support for 'longer or shorter' ACS transcripts in other plant species as can be determined by BLAST analyses of sequence database. Regardless

of the start site, BLAST alignment between the plant and ACS proteins from other species indicates that the aa identity resides within aa 91-to 743 of Arabidopsis.

SDS-PAGE analysis (Figure S1, panel A) of total protein isolated from *E. coli* cells expressing $ACS_{97\text{-}743}$ (Line 2) showed a distinct 75 kDa protein, corresponding to the calculated molecular weight of the recombinant ACS. The ACS was subsequently purified by a gel-filtration (Superdex 75) chromatography followed by Q-sepharose chromatography (lane 3). The resulting ACS protein band was distinct and was not observed in control protein fractions, isolated from *E. coli* cells expressing an empty vector, and purified by the same procedures (lane 4). The authenticity of the ACS was determined by cutting the 75 kDa band from the gel, digesting with trypsin, and confirming its aa identity by MALDI-TOF.

To determine ACS activity, enzymatic reactions were resolved by ion-paired reverse phase HPLC. Figure S1(B) (panel 4) shows that, in the presence of acetate, Mg^{2+} and ATP, recombinant ACS readily converts CoA to AcCoA. In addition to a peak migrating as AcCoA [17.2 min, Figure S1(B)] a specific peak eluting as AMP (8.5 min) was observed, while the ATP peak (11.5 min) was reduced in size. Control protein, (i.e. protein extracted from *E. coli* expressing an empty vector and purified following the same procedure) was unable to produce AcCoA [Figure S1(B), panel 5]. This result indicated that the recombinant ACS is an active Acetyl CoA Synthase. Interestingly, both AMP and ADP were identified as final products, however ADP was also 'made' in the negative control, and we attribute this to the instability of ATP. Nevertheless, to ascertain whether ACS is an AMP-forming or ADP-forming enzyme, the reverse reactions were carried out. ACS was incubated with AcCoA, PPi and AMP, or with AcCoA, phosphate and ADP. The resulting data indicate that ACS was able to convert AcCoA to CoA only in the presence of AMP [Figure S1(B), panel 6]. The enzyme was unable to convert ADP, Pi and AcCoA to a product [Figure S1(B), panel 8], indicating that the plant ACS, like the bacteria and human enzyme, is a truly AMP-forming enzyme. To validate that the HPLC peak eluted at 17.2 min is AcCoA, the peak marked 2 [Figure S1(B), panel 4] was collected and its identity was confirmed by ^{1}H NMR spectroscopy as AcCoA [Figure S1(C)].

ACS had high enzymatic activity at pH 7.6, and activity was reduced by 25% and 45% at pH 7.0 and pH 6.5, respectively (Table S1). The enzyme requires metals (Mg^{2+}) and no activity was observed in the presence of EDTA. Interestingly, ACS displayed similar relative activity if Mg was substituted by Ca^{2+}. However, activity was lost when Mg^{2+} in the reaction was substituted by Mn^{2+}, Zn^{2+}, Cu^{2+} or Fe^{2+}. ACS had high enzymatic activity between 25 to 37ºC and activity was completely inhibited above 50ºC. To test the substrate specificity of ACS, propionate, malonate, butyrate and succinate were examined and compared with acetate. Relative to acetate (Table S1), the recombinant Arabidopsis ACS converted ~40% propionate to propioanate-CoA, like the bacterial (*6*) and human ACS, but conversion of other non-acetyl substrates to CoA derivatives was not observed. For the kinetic studies of ACS, various concentrations of ATP were tested while the other ACS substrates (acetate, and CoA), were kept at saturating levels. Similarly, concentrations of CoA were varied while acetate and ATP

concentrations were kept at saturating levels. The apparent K_m values for ATP, acetate, and CoA were 75.9 μM, 139 μM, 115 μM respectively and the Vmax values for these three substrates were 1164 min^{-1}, 1248 min^{-1}, and 1600 min^{-1}, respectively. In contrast to the ACS activity isolated from spinach leaf that absolutely required DTT (*7*), the activity of recombinant *Arabidopsis* ACS is independent of DTT. Moreover, enzymatic activity of ACS was inhibited by 95% in the presence of 10 mM DTT.

Supplementary References:

1. Gu, X., and Bar-Peled, M. (2004) The biosynthesis of UDP-galacturonic acid in plants. Functional cloning and characterization of Arabidopsis UDP-D-glucuronic acid 4-epimerase, *Plant Physiol 136*, 4256-4264.
2. Devarajan, K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology, *PLoS Comput Biol 4*, e1000029.
3. Lee, D. D., and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization, *Nature 401*, 788-791.
4. Battogtokh, D., Asch, D. K., Case, M. E., Arnold, J., and Schuttler, H. B. (2002) An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa, *Proc Natl Acad Sci U S A 99*, 16904-16909.
5. Yu, Y., Dong, W., Altimus, C., Tang, X., Griffith, J., Morello, M., Dudek, L., Arnold, J., and Schuttler, H. B. (2007) A genetic network for the clock of Neurospora crassa, *Proc Natl Acad Sci U S A 104*, 2809-2814.
6. Luong, A., Hannah, V. C., Brown, M. S., and Goldstein, J. L. (2000) Molecular characterization of human acetyl-CoA synthetase, an enzyme regulated by sterol regulatory element-binding proteins, *J Biol Chem 275*, 26458-26466.
7. Zeiher, C. A., and Randall, D. D. (1991) Spinach Leaf Acetyl-Coenzyme A Synthetase: Purification and Characterization, *Plant Physiol 96*, 382-389.
8. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.
9. Schwarz, Gideon E. (1978) Estimating the dimension of a model, *Annals of Statistics 6*, 461–464.
10. W. Dong, X. Tang, Y. Yu, R. Nilsen, R. Kim, J. Griffith, J. Arnold, and H.-B. Schüttler (2008) Systems Biology of the Clock in Neurospora crassa, *PLoS ONE **3** (8): e3105* (28 pages).

Table S1: Summary of enzymatic properties of Arabidopsis ACS. The activity of ACS in a standard reaction system as described in Materials and Methods was set to 100% for this comparative studies. n.d.(not detected). The data is an average of three repeats.

		Relative ACS activity
Substrate specificity	Acetate (1mM)	100%
	Propionate (1mM)	45%
	Malobate (1mM)	n.d.
	Butyrate (1mM)	n.d.
	Succinate (1mM)	n.d.
pH	6.5	63%
	7.0	87%
	7.5	100%

	8.0	92%
	9.0	81%
Temperature (°C)	0	n.d.
	25	90%
	30	97%
	37	100%
	50	n.d.
Metal ions (5 mM)	Mg^{2+}	100%
	Ca^{2+}	100%
	Mn^{2+}	n.d.
	Zn^{2+}	n.d.
	Fe^{3+}	n.d.
	Cu^{2+}	n.d.
Additives	DTT (10 mM)	4.3%
	NaCl (100 mM)	50%

Table S2: Rate coefficient ensemble averages and ensemble standard deviations are given in the 3rd and 4th column, respectively, in units of $(mM)^{(1-n)}$/minute for reaction of order *n* (*i.e.,* a reaction having *n* reactants entering), abbreviation: ATP (T), acetate(A), CoA (C), acetyl-AMP (Q), AMP (M), AcCoA (R). Forward reactions listed without a corresponding backward reaction are assumed to be irreversible, *i.e.*, have been assigned a fixed, zero backward rate coefficient.

Model	Reaction Equation	Rate Coefficient Ensemble Average	Rate Coefficient Ensemble SD
1	ATP+acetate+CoA+ACS-->E/A/T/C	4.52E+03	1.52E+04
	E/A/T/C -->ATP+acetate+CoA+ACS	4.62E+02	1.67E+03
	E/A/T/C-->E/R/M/P	1.04E+03	1.63E+03
	E/R/M/P-->E/A/T/C	4.46E+02	1.57E+03
	E/R/M/P-->AcCoA+AMP+PPi+ACS	1.51E+03	2.43E+03
	AcCoA+AMP+PPi+ACS -->E/R/M/P	5.94E+04	1.32E+05
	ATP-->ADP	2.66E-03	3.67E-05
2	ATP+acetate+ACS-->E/A/T	2.45E+02	3.36E+02
	E/A/T --> ATP+acetate+ACS	4.70E+02	4.89E+02
	E/Q/P-->Acetyl-AMP+PPi+ACS	1.91E+03	1.69E+03
	Acetyl-AMP+PPi+ACS --> E/Q/P	9.41E+02	1.04E+03
	E/A/T-->E/Q/P	1.46E+03	1.35E+03
	E/Q/P --> E/A/T	6.78E+04	1.81E+05
	Acetyl-AMP+CoA+ACS-->E/Q/C	3.50E+03	6.63E+03
	E/Q/C --> Acetyl-AMP+CoA+ACS	4.12E+02	9.48E+02
	E/Q/C-->E/R/M	1.48E+03	1.35E+03
	E/R/M --> E/Q/C	2.22E+02	6.81E+02
	E/R/M-->AcCoA+AMP+ACS	9.63E+02	1.34E+03
	AcCoA+AMP+ACS --> E/R/M	2.38E+03	2.36E+03
	ATP-->ADP	3.32E-03	3.68E-04
3	ATP+acetate+ACS-->E/A/T	3.32E+02	3.88E+02
	E/A/T --> ATP+acetate+ACS	9.33E+02	1.70E+03
	E/A/T-->E/Q/P	8.36E+02	6.92E+02
	E/Q/P --> E/A/T	5.34E+01	1.15E+02
	E/Q/P-->Acetyl-AMP+PPi+ACS	1.64E+03	2.28E+03
	Acetyl-AMP+PPi+ACS --> E/Q/P	1.79E+05	2.25E+05
	E/Q/P+CoA-->E/Q/C+PPi	8.56E+02	1.62E+03
	E/Q/C+PPi --> E/Q/P+CoA	1.56E+03	3.25E+03
	E/Q/C-->E/R/M	1.77E+03	2.25E+03
	E/R/M --> E/Q/C	7.04E+02	1.96E+03
	E/R/M-->AcCoA+AMP+ACS	8.63E+02	1.85E+03
	AcCoA+AMP+ACS --> E/R/M	1.40E+04	4.32E+04
	ATP-->ADP	3.22E-03	3.97E-04
4	ATP+acetate+ACS-->E/A/T	9.14E+02	1.72E+03
	E/A/T --> ATP+acetate+ACS	5.26E+02	7.13E+02
	E/A/T-->E/Q/P	1.28E+03	1.82E+03
	E/Q/P --> E/A/T	1.10E+03	1.64E+03
	E/Q/P-->E/Q+PPi	1.06E+03	1.20E+03
	E/Q+PPi --> E/Q/P	6.63E+02	9.95E+02
	E/Q+CoA-->E/Q/C	2.02E+03	4.08E+03
	E/Q/C --> E/Q+CoA	1.46E+02	3.38E+02
	E/Q/C-->E/R/M	2.11E+03	2.42E+03
	E/R/M --> E/Q/C	4.35E+02	7.58E+02
	E/R/M-->AcCoA+AMP+ACS	1.62E+03	1.76E+03
	AcCoA+AMP+ACS --> E/R/M	3.07E+04	1.03E+05
	E/Q-->Acetyl-AMP+ACS	1.28E+03	1.57E+03
	Acetyl-AMP+ACS --> E/Q	6.04E+04	1.52E+05
	ATP-->ADP	3.15E-03	3.75E-04

Figure Legends

Figure S1. Isolation and biochemical characterization of recombinant ACS. (A) SDS-PAGE analyses of total *E. coli* protein isolated from cell expressing ACS plasmid (lane 1), gel filtration followed by Source-Q column purification of ACS (lane 3), and control cells expressing empty plasmid (lane 2) or column purified control (Lane 4). The protein band (~ 75 kDa) points by an arrow, was excised from the gel, trypsin digested and analyzed by MALDI-TOF; and confirmed the aa identity of the recombinant Arabidopsis ACS. (B) Forward and reverse HPLC-based ACS assays. Forward ACS activity was measured with purified ACS (panel 4) or control protein fraction isolated and purified as ACS but from cells expressing empty plasmid (panel 5) in the presence of 1 mM ATP, 1 mM acetate and 0.2 mM CoA. Note the formation of two new peaks migrated with retention time as std AMP (peak # 5) and AcCoA (peak #2) and reduction of CoA (peak #1). Peak x is unknown contaminate in the CoA preparation. Formation of ADP peak (peak #4) is likely an artifact due to instability of ATP (peak #3), and observed in negative control as well. The reverse ACS activity was measured with purified recombinant ACS in the presence of 1 mM of AMP, 1 mM of PPi and 1 mM of AcCoA. The formation of two new peaks migrated with retention time as std ATP and CoA was observed (panel 5), while no similar peaks was observed in the presence of 1 mM of ADP, 1 mM of Pi and 1 mM of AcCoA (panel 7). (C) Peak#2 in panel 4 was collected; lyophilized, analyzed by 1H-NMR and confirmed as AcCoA when compared with Std.

Figure S2. Network reconstruction of AtACS-catalyzed reaction. (A) Time course of concentrations of CoA, AcCoA, acetate and PPi to show how the ENS data fit the MPSF experimental data for different models. The scatter dot represents the experimental data. The solid line represents the ENS data (middle line) with standard derivation (upper and lower lines). (B) The χ^2-values of ENS MC random walking process PIBS data sets of all models. The first 10000 for PIBS walking points were shown and this process was repeated 10 times.

Figure S3. Time course of residuals of AMP for the ENS fit to MPSF-processed experimental data for models 1, 2, 3 and 4 in panels (A), (B), (C), and (D), respectively. Residuals were calculated as the difference between re-scaled experimental concentration data $\exp(Y_j-\langle\phi_{c(j)}\rangle)$ and the predicted concentration $\exp(\langle F_j(\theta)\rangle)$, obtained from the ENS mean of the natural log concentration $\langle F_j(\theta)\rangle$.

Figure S1

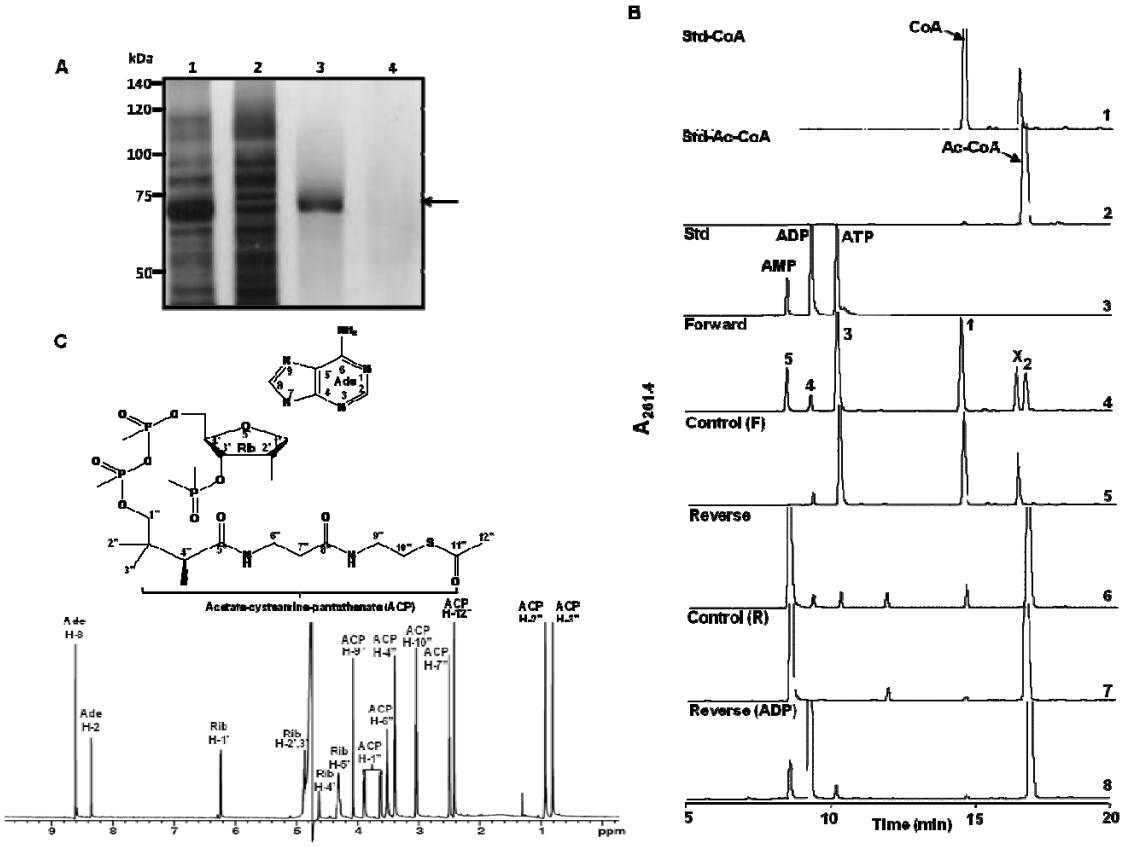

A
kDa
1
2
3
4
140
120
100
75
50
B
Std-CoA
CoA
Std-Ac-CoA
Ac-CoA
Std
ADP
ATP
AMP
Forward
Control (F)
Reverse
Control (R)
Reverse (ADP)
$A_{261.4}$
Time (min)
C
Acetate-cysteamine-pantothenate (ACP)
ppm

Figure S2

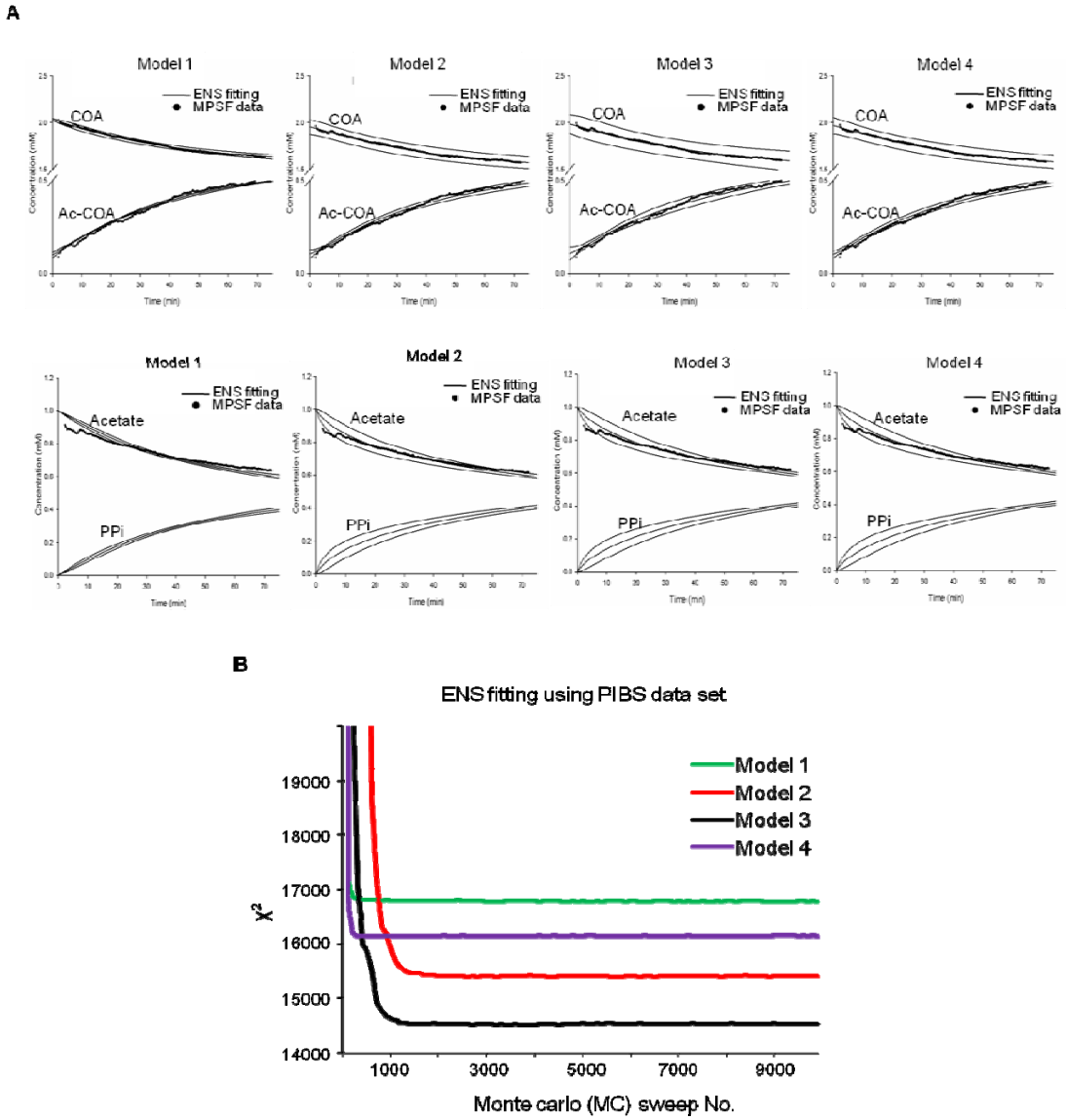
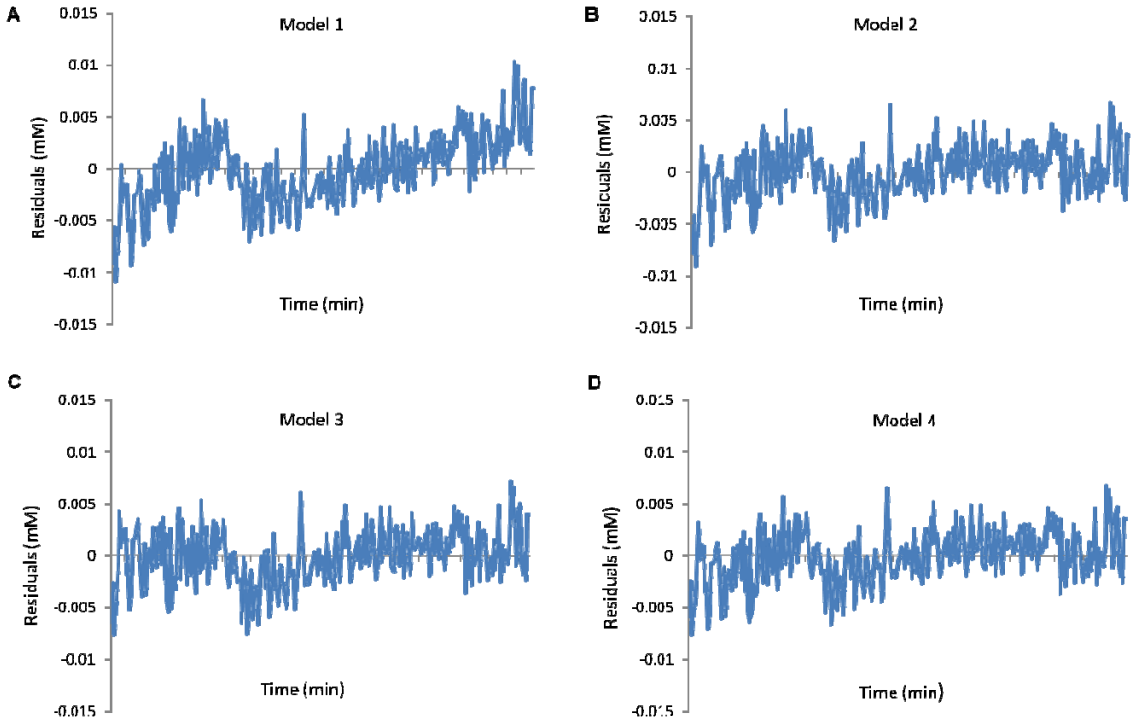

A
Model 1
Model 2
Model 3
Model 4
ENS fitting
MPSF data
COA
Ac-COA
Acetate
PPi
Time (min)
B
ENS fitting using PIBS data set
Model 1
Model 2
Model 3
Model 4
19000
18000
17000
16000
15000
14000
1000
3000
5000
7000
9000
Monte carlo (MC) sweep No.

Figure S3


A
Model 1
Residuals (mM)
Time (min)
B
Model 2
Residuals (mM)
Time (min)
C
Model 3
Residuals (mM)
Time (min)
D
Model 4
Residuals (mM)
Time (min)