Analytical Biochemistry **285,** 33–49 (2000)
doi:10.1006/abio.2000.4744, available online at http://www.idealibrary.com on IDEAL®

# A Holistic Approach for Protein Secondary Structure Estimation from Infrared Spectra in $H_2O$ Solutions

Ganesh Vedantham,* H. Gerald Sparks,†,[1] Samir U. Sane,†,[2] Stelios Tzannis,†,[3] and Todd M. Przybycien*,[4]

*Applied Biophysics Laboratory, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; and †Howard P. Isermann Department of Chemical Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180*

Received January 31, 2000

**We present an improved technique for estimating protein secondary structure content from amide I and amide III band infrared spectra. This technique combines the superposition of reference spectra of pure secondary structure elements with simultaneous aromatic side chain, water vapor, and solvent background subtraction. Previous attempts to generate structural reference spectra from a basis set of reference protein spectra have had limited success because of inaccuracies arising from sequential background subtractions and spectral normalization, arbitrary spectral band truncation, and attempted resolution of spectroscopically degenerate structure classes. We eliminated these inaccuracies by defining a single mathematical function for protein spectra, permitting all subtractions, normalizations, and amide band deconvolution steps to be performed simultaneously using a single optimization algorithm. This approach circumvents many of the problems associated with the sequential nature of previous methods, especially with regard to removing the subjectivity involved in each processing step. A key element of this technique was the calculation of reference spectra for ordered helix, unordered helix, sheet, turns, and unordered structures from a basis set of spectra of well-characterized proteins. Structural reference spectra were generated in the amide I and amide III bands, both of which have been shown to be sensitive to protein secondary structure content. We accurately account for overlaps between amide and nonamide regions and allow different structure types to have different extinction coefficients. The agreement between our structure estimates, for proteins both inside and outside the basis set, and the corresponding determinations from X-ray crystallography is good.** 

***Key Words:*** **infrared spectroscopy; spectral deconvolution; protein secondary structure; reference spectra.**

Fourier-transformed infrared (FTIR)[5] spectroscopy is perhaps the most versatile spectroscopic technique for analyzing protein secondary structure in diverse physiochemical environments. FTIR spectroscopy has been applied to investigate protein structure in solution (1, 2), in aggregates and inclusion bodies (3, 4), as well as during lyophilization (5–7) and freeze/thaw processing (8). In addition, attenuated total reflection (ATR) FTIR spectroscopy is ideal for studying protein adsorption onto catheter surfaces (9), chromatographic media (10–12), and a variety of other polymeric surfaces (13–15).

In the past decade, a plethora of methods to estimate protein secondary structure contents via analysis of amide I, II, and III, band spectra have been reported. These methods include, but are not limited to, solitary use or combinations of factor analysis (FA) (16–20), singular value decomposition (SVD) (21, 22), Fourier self-deconvolution (FSD; or resolution enhancement) (14, 23–26), second derivative (SD) band identification and fitting (27–29), and the development of spectral correlation coefficients (30, 31). Recent reviews of these

[1] Current address: DuPont Experimental Station, Route 141 and Henry Clay Road, Wilmington, DE 19880.

[2] Current address: Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080.

[3] Current address: Inhale Therapeutics, 150 Industrial Road, San Carlos, CA 94070.

[4] To whom correspondence should be addressed. Fax: (412) 268 7139. E-mail: todd@andrew.cmu.edu

[5] Abbreviations used: FTIR, Fourier-transformed infrared; ATR, attenuated total reflection; FA, factor analysis; SVD, singular value decomposition; FSD, Fourier self-deconvolution; SD, second derivative; R.H.S., right-hand side; GL, Gaussian–Lorentzian.

0003-2697/00 $35.00

techniques by Pelton and McLean (32) and Jackson and Mantsch (33) are instructive. This large body of work devoted to protein secondary structure estimation from infrared spectra has led to a number of discrepancies that persist throughout the literature.

In a classic work cited by virtually every researcher in the field, Byler and Susi (24) used FSD to analyze the spectra of 21 globular proteins in $^2H_2O$ and were able to assign components of amide I band spectra to helices, $\beta$-sheet, turns, and random (unordered) structure. By their method, segments with similar structure do not necessarily exhibit peaks with identical frequencies from protein to protein. For example, Byler and Susi (24) reported frequencies varying from 1651 to 1657 $cm^{-1}$ for helical vibrations in proteins, and frequencies for homopolypeptides in helical conformations have been reported as low as 1634 $cm^{-1}$ (24). Also, Chirgadze *et al.* (34) reported that for helical structures, the corresponding peak width increases with decreasing helical order. In light of this, when deconvoluting protein amide bands, many algorithms involve subjective peak assignments or allow the peak positions and widths to vary during the structure estimation procedure. To circumvent these difficulties, many authors have invoked either resolution enhancement or second derivative techniques to help identify the positions of relevant peaks, followed by the assignment of a structure type to each peak and a fit of each peak with Gaussian and/or Lorentzian distribution functions. However, significant bias in the results can still be introduced because choices of the resolution enhancement factor for FSD and the peak assignments in both methods are subjective.

An alternative to case-specific peak assignment methods is the direct or indirect development of structural reference spectra, or eigenspectra, that theoretically represent either pure motifs, such as $\alpha$-helix, $\beta$-sheet, and turns, or linear combinations of pure motifs (33). These idealized spectra are then fit to the spectrum of a protein of unknown structure by varying the corresponding motif fractions. These fractions serve as the weighting factors in a linear superposition scheme. The reference spectra are generated by the decomposition of a calibration set or basis set of real protein spectra covering a broad range of structural fractions, utilizing methods such as SVD, band fitting, or matrix inversion (17, 19, 21, 35). The reference spectra approach has been successfully applied to both CD spectra (36) and Raman spectra (37), but has had mixed results when applied to FTIR spectra (19, 21). Contrary to the results of Byler and Susi (24), the reference spectra method assigns fixed positions to peaks representing the various structure motifs. However, as will be demonstrated by the results of this work, the mixed success of the past reference spectra methods for protein secondary structure predictions from FTIR spectra is associated with the structure class assignments and not the seeming contradiction with the work of Byler and Susi (24).

In addition to uncertainty in peak positions and assignments, the shortcomings of most previous routines involve the sequential subtraction of background solvent and water vapor contributions to the protein solution spectra, followed by an arbitrary baseline assignment to isolate the amide region of interest. Baseline correction can be a function of operator experience with the subtraction procedure (38). Obtaining a so-called "flat-region" in the 1750–2200 $cm^{-1}$ frequency range is the typical criterion used for bulk water subtraction. The degree of background subtraction is often determined manually and "flat" is rarely quantified. After all background subtractions, the amide I region is often isolated for analysis by truncating the spectrum at 1600 and 1700 $cm^{-1}$, followed by the subtraction of a linear baseline to zero the ends of the spectrum. When examining the amide I and II regions together, end points of 1480 and 1700 $cm^{-1}$ are typically used, while the amide III region is often bounded at 1200 and 1300 $cm^{-1}$ (39). In this subjective approach, an early error in sequential background and baseline subtractions will be carried through to the band fitting or reference spectra routine and will produce potentially erroneous results. Additionally, choosing arbitrary end points for a baseline subtraction ignores any contributions from adjacent vibrational modes that tail into the amide regions and vice versa.

No current algorithm for protein secondary structure estimation from infrared spectra accounts for the impact of solutes on background solvent spectra or the possibility that different secondary structure motifs may absorb with varying extinction coefficients. As demonstrated via Raman spectroscopy, the O–H bending and stretching vibrations of water undergo significant changes in the presence of proteins and other solutes (37, 40, 41). Increasing evidence also supports the idea that different molar extinction coefficients exist for the various structure types contributing to the protein amide vibrations (33, 42, 43). Accurate subtraction of background solvent and assignment of the proper weights to the amide band components are critical for obtaining reliable secondary structure estimates, especially in cases involving low protein concentrations.

Another major discrepancy in current protein structure estimation algorithms concerns the paradox seemingly generated when normalizing spectra. It is common practice during analysis to normalize a spectrum after all background subtractions have been performed and a particular amide band has been isolated. However, to accurately account for all the overlapping regions between peaks that correlate with protein structure and those that do not, the amide region should be normalized before subtraction. In addition, possible

**TABLE 1**

List of Proteins Used for FTIR Spectroscopic Studies

| Abbreviation | Protein | Source | Cat. No. | Lot No. | PDB file[b] |
|---|---|---|---|---|---|
| ALA | α-Lactalbumin | Bovine milk | L-5385 | 92H7015 | 1hfx |
| BGH | Bovine growth hormone | *E. coli* (recombinant) | | M901-004 | 1bst |
| BLB | β-Lactoglobulin | Bovine milk | L-8005 | 13H7150 | 1beb |
| CAL | Conalbumin | Chicken egg white | C-0755 | 116H7035 | 1aiv |
| CAN | Carbonic anhydrase | Bovine erythrocytes | C-3934 | 47H1358 | 2cba |
| CHY[a] | α-Chymotrypsin | Bovine pancreas | C-7762 | 27H7010 | 5cha |
| CON[a] | Concanavalin A | Canavalia ensiformis | C-7275 | 118F7160 | 1apn |
| CYT | Cytochrome *c* | Horse heart | C-7752 | 25H7045 | 1hrc |
| HSA | Human serum albumin | Human serum | A-9511 | 24H9314 | 1bj5 |
| LYS[a] | Lysozyme | Chicken egg white | L-6876 | 65H7025 | 1azf |
| MYO[a] | Myoglobin | Sperm whale | M-7527 | 17H6660 | 104m |
| PAP[a] | Papain | Papaya latex | P-4762 | 107H7015 | 1ppn |
| PEP | Pepsin | Porcine stomach mucosa | P-6887 | 120H8095 | 4pep |
| RNA[a] | Rnase | Bovine pancreas | R-5500 | 86H7046 | 3rn3 |
| SUB[a] | Subtilisin-BPN′ | Bacillus licheniformis | 101129 | 69618 | 2st1 |
| TPI[a] | Triosephosphate isomerase | Rabbit muscle | T-6258 | 96H9554 | 1ag1 |

[a] Included in the basis set for generation of the reference spectra.
[b] Protein Data Bank file listing. URL: http://www.rcsb.org/pdb/.

variations in secondary structure extinction coefficients imply that the areas of the amide bands also depend on the overall protein secondary structure content. This enigma can be resolved by performing the subtractions, normalization, and deconvolution of the amide band of interest simultaneously.

In this paper, we describe a holistic reference spectra calculation technique for the generation of idealized reference infrared spectra in the amide I and amide III regions, followed by a procedure for the estimation of protein secondary structure for unknown samples. Our prediction technique did not make use of the amide II region because this vibrational mode has been shown to be less sensitive to variations in protein secondary structure content (39). In the calculation of the reference spectra, all subtractions, normalization, and amide band deconvolution steps are performed simultaneously, following the method Sane and co-workers (37) developed for Raman spectral deconvolution. All non-structure-related vibrational peaks are fit using equally weighted Gaussian–Lorentzian product functions; peaks correlating with protein secondary structure are allowed to have different molar extinctions. This method places no restrictions on the frequency ranges analyzed: overlaps between non-structure- and structure-associated peaks are accounted for since all components are fit simultaneously. The introduction of a protein-dependent effective concentration variable solved the normalization problem. The calculation of reference spectra involved multivariate nonlinear least-squares minimization which was implemented in Matlab 5.0 (Mathworks Inc., Natik, MA). The idealized reference spectra were optimized for internal consistency via a bootstrapping algorithm. FTIR spectra of proteins outside the basis protein set were then analyzed to validate the secondary structure estimation algorithm. Results presented here for calculated structural reference spectra compare well with those in the literature and provide good secondary structure estimates for proteins.

## MATERIALS AND METHODS

### *Materials*

The proteins in and outside the reference set were chosen to cover a broad range of secondary structure motifs; a list of the proteins studied is given in Table 1. The protein's secondary structure assignment is dependent on the choice of assignment algorithm (44, 45). In this report, all secondary structure assignments were made using the STRIDE algorithm of Frishman and Argos (45). The use of a single assignment algorithm eliminates the discrepancies that ensue from the application of dissimilar criteria and algorithms to crystallographic data (46). The STRIDE secondary structure assignments of the proteins analyzed in this report, both within and outside the reference set, are shown on a triangular diagram in Fig. 1. We have assigned STRIDE-identified $3_{10}$ helices as well as α-helices of three or less contiguous residues as unordered helices in this work. The Protein Data Bank files used to generate the STRIDE estimates are listed in Table 1. All the proteins studied exhibit significant ordered secondary structure content in their native states.

Subtilisin BPN′ was purchased from ICN Biomedicals Inc. (Irvine, CA). Bovine growth hormone was a gift from Monsanto (St. Louis, MO). All other proteins, see Table 1 for abbreviations used throughout this work, and reagents for buffers were purchased from Sigma Chemical

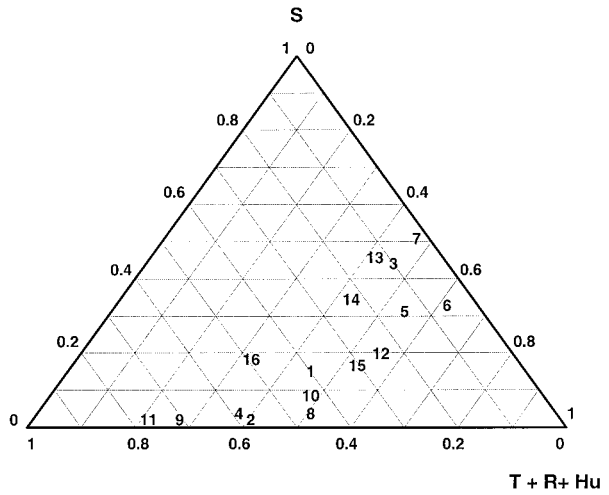


**FIG. 1.** Secondary structure assignments for proteins analyzed in this work: 1, ALA; 2, BGH; 3, BLB; 4, CAL; 5, CAN; 6, CHY; 7, CON; 8, CYT; 9, HAS; 10, LYZ; 11, MYO; 12, PAP; 13, PEP; 14, RNA; 15, SUB; 16, TPI. All structure assignments were based on the STRIDE algorithm of Frishman and Argos (45). Symbols: S, total sheet; Ho, ordered helix; T + R + Hu, turn + random coil + unordered helix.

Co. (St. Louis, MO). The final buffer conditions used for all protein solutions are listed in Table 2. Several proteins required processing to remove additives. Lysozyme and papain were dissolved into their respective buffers and then dialyzed with Spectra/Por Biotech 500 MWCO cellulose ester membranes (Cat. No. 08-750-1A), purchased from Fisher Scientific Inc. (Pittsburgh, PA), to remove sodium acetate; dialyses were conducted against 500-mL reservoirs of final buffer solutions for 12 h, with three intermediate reservoir changes. In addition, solids remaining in the lysozyme solution were sedimented in a Eppendorff 5415C microcentrifuge (Brinkman Instruments, Westbury, NY) at 14,000 rpm for 15 min and the supernatant was pipetted off for study. Triose phosphate isomerase was dialyzed, as described above, to remove borate and EDTA. Myoglobin was obtained in liquid form at a concentration of 4.8 mg/mL. All other proteins were dissolved directly into the corresponding buffers listed in Table 2. $\alpha$-Chymotrypsin was centrifuged, as above, to remove residual solids. After dissolution, the concanavalin A protein solution remained slightly cloudy; however, centrifugation precipitated the protein and thus the turbid solution was used for analysis. In addition, myoglobin and papain were concentrated in a Beckmann Instruments, Inc. (Palo Alto, CA), TJ-6 centrifuge at 3000 rpm to a final volume of 250 $\mu$L with Centricon-3, 3000 MWCO, centrifugal membrane concentrators from Amicon, Inc. (Beverly, MA). Prior to protein dissolution, all buffers were filtered through syringe filters with 0.45-$\mu$m nylon membranes to remove dust and undissolved salts. The proteins included in the reference set are CHY, CON, LYS, MYO, PAP, RNA, SUB, and TPI.

### *FTIR Spectroscopy*

All protein spectra were recorded in $H_2O$ solution. All spectra were collected with a Nicolet Magna 550 Series II FTIR spectrometer (Madison, WI) with a horizontal ATR accessory from SpectraTech, Inc. (Shelton, CT). The ATR accessory used a trapezoidal germanium crystal (7.0 × 1.0 cm; length × width), with ends cut to 45° generating 12 internal reflections, that was mounted into a sample-

**TABLE 2**

Protein Solution Conditions and Spectral Quality

| Protein | Buffer | Protein concentration (mg/ml) | S/N (amide I band) | S/N (amide III band) |
|---|---|---|---|---|
| ALA | 10 mM sodium phosphate, pH 6.0, with 100 mM NaCl | 28 | 209 | 22 |
| BGH | DI with a trace of HCl, pH 3.8[a] | 8 | 178 | 33 |
| BLB | 50 mM sodium phosphate, pH 7.0 | 40 | 161 | 65 |
| CAL | 100 mM NaCl, pH 6.0 | 20 | 167 | 103 |
| CAN | DI water | 24 | 669 | 74 |
| CHY | DI with a trace of HCl, pH 3.8[a] | 18 | 184 | 42 |
| CON | DI with a trace of HCl, pH 3.8[a] | 21 | 439 | 38 |
| CYT | 25 mM sodium phosphate, pH 6.0, with 100 mM NaCl | 15 | 186 | 35 |
| HSA | 25 mM sodium phosphate, pH 7.0, with 100 mM NaCl | 28 | 225 | 167 |
| LYS | DI water | 32 | 877 | 73 |
| MYO | 20 mM Tris–HCl, pH 8.0 | 18 | 172 | 25 |
| PAP | DI with a trace of HCl, pH 3.8[a] | 27 | 195 | 27 |
| PEP | 25 mM sodium phosphate, pH 7.0 with 100 mM NaCl | 22 | 172 | 160 |
| RNA | DI with a trace of HCl, pH 3.8[a] | 18 | 402 | 52 |
| SUB | 25 mM sodium phosphate, pH 6.0, with 100 mM NaCl | 19 | 389 | 48 |
| TPI | DI water | 18 | 211 | 13 |

[a] Trace is defined as approximately 50 to 100 $\mu$l of 2 M HCl in 1 liter DI water.

boat/trough. The spectrometer was equipped with a liquid nitrogen-cooled mercury cadmium telluride detector. To reduce the contributions of water vapor and carbon dioxide, the IR system was continuously purged with air from a Balston, Inc. (Haverhill, MA) 75-45 FTIR Purge Gas Generator at 30 standard cubic feet per minute and supplemented with nitrogen gas from the vent of a liquid nitrogen tank. To obtain protein solution and corresponding buffer background spectra, approximately 250 $\mu$L of each solution was spread evenly to completely cover the germanium crystal. The crystal was then sealed with parafilm to minimize evaporation during acquisition. Protein concentrations above 20 mg/mL ensured that less than 2% of the FTIR signal derived from molecules adsorbed to the germanium crystal, assuming a worst case scenario of monolayer coverage attained by random sequential adsorption with a jamming limit of 55%. All ATR-corrected spectra were collected in the 1000 to 4000 $cm^{-1}$ range as sets of 2048 time-averaged, double-sided interferograms with Happ–Genzel apodization. Spectral resolution was set at 2 $cm^{-1}$ and a gain of 8 and an aperture of 38 were used. After each experiment, the exposed surface of the germanium crystal was cleaned via a five-step process: (1) rinsing with DI water, (2) soaking in a 1% (w/w) SDS solution for 10 min, (3) rinsing thoroughly with DI water, (4) rinsing thoroughly with a 50% (w/w) aqueous ethanol solution, and (5) drying with compressed air filtered through cotton to remove oils and particulates. Amide I band signal-to-noise (S/N) ratios varied from 877 to 161, whereas amide III band S/N ratios varied from 166 to 12, as shown in Table 2. Amide band S/N ratios were calculated as 2.5 times the maximum intensity of the background-subtracted band divided by 3 times the standard deviation of the intensity between 1850 and 2200 $cm^{-1}$.

### *Data Analysis*

*Mathematical representation of protein FTIR spectra.* In addition to the secondary structure-sensitive amide I and amide III bands, there are several other vibrational modes active in the spectral region of interest, including the amide II band. Protein solution FTIR spectra also contain background contributions from buffer and water vapor. In addition, spectra may have a contribution from a sloping baseline. By assuming that the contributions of all underlying spectral components are additive, invoking the principle of superposition, any set of spectra from $p$ proteins ($p > 1$) can be represented in matrix form as

$$I^{\text{calc}}_{z\times p} = v_{z\times 1}a_{1\times p} + 1_{z\times 1}b_{1\times p} + B_{z\times m}A_{m\times p} + N_{z\times n}D_{n\times p} + [S^{\text{I}}_{z\times q}E^{\text{I}}_{q\times r}F^{\text{I}}_{r\times p} + S^{\text{III}}_{z\times s}E^{\text{III}}_{s\times t}F^{\text{III}}_{t\times p}]C^{\text{eff}}_{p\times p}, \quad [1]$$

where $I^{\text{calc}}_{z\times p}$ is the calculated spectral intensity for $p$ proteins at $z$ frequencies. All subscripts in Eq. [1] correspond to the dimensions (rows × columns) of the associated matrices, each of which will be elaborated upon below.

The first two terms on the right-hand side (R.H.S.) of Eq. [1] describe a linear baseline for the spectral range of interest, 1000 to 2200 $cm^{-1}$, during the optimization routine. Here $v_{z\times 1}$ and $1_{z\times 1}$ are vectors of length $z$ containing frequencies and ones, respectively. The baseline slope and intercept for each protein spectrum are compiled in the vectors $a_{1\times p}$ and $b_{1\times p}$, respectively.

Background contributions from buffer (or solvent; $m = 1$), water vapor ($m = 2$), and, where necessary, an underlying surface ($m = 3$), are accounted for in the third term on the R.H.S. of Eq. [1]. The matrix $B_{z\times m}$, representing $m$ independently measured background spectra recorded at $z$ frequencies, is multiplied by the matrix of background signal magnitudes (or amplitudes), $A_{m\times p}$, containing the respective background contributions to each protein spectrum.

The fourth term on the R.H.S. of Eq. [1] accounts for the vibrational peaks in the frequency range analyzed that are not correlated with protein secondary structure, here on designated as nonstructure peaks. These peaks embody vibrations associated with amino acid side chains and the amide II band. We have not included individual side-chain resonances that contribute intensity in the amide I and III bands (47). These resonances typically account for 5 to 15% of the signal intensity in the amide I region (43), but are highly variable in position from protein to protein (33).

Each individual peak $i$ is expressed as a Gaussian–Lorentzian (GL) product function

$$\text{GL}_i = \left[\frac{2pw_i}{\pi(pw_i^2 + 4(\bar{\nu}_i - \nu))^2}\right]^Y \times \left[\frac{2\sqrt{\frac{\ln(2)}{\pi}}}{pw_i}\exp\left\{\frac{-4\ln(2)(\bar{\nu}_i - \nu)^2}{pw_i^2}\right\}\right]^{(1-Y)} \quad [2]$$

each of which has an associated mean frequency position, $\bar{\nu}_i$, and peak width at half-height, $pw_i$. Equation [2] is used to generate $n$ nonstructure peaks at $z$ frequencies across the whole spectral range, forming the matrix $N_{z\times n}$. The matrix $D_{n\times p}$ contains the nonstructure peak magnitudes (or amplitudes) for each corresponding protein in the reference set. In our formulation, the number, associated mean peak positions, and peak widths of nonstructure GL peaks are identical for each protein (i.e., protein independent); however, the amplitudes corresponding to the contribution of each nonstructure peak to an individual protein spectrum vary from protein to protein. The exponent $Y$ in Eq. [2] is a weighting factor that determines the relative Gaussian–Lorentzian character of the nonstructure

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.