Bile Acid Chemistry, Biology, and Therapeutics

During the Last 80 Years: Historical Aspects

Alan F. Hofmann and Lee R. Hagey

Running footnote: History of bile acid research

Department of Medicine, University of California, San Diego 92093-063

Correspondence to Alan F. Hofmann (AHofmann@ucsd.edu) or Lee R. Hagey (LHagey@ucsd.edu)

Alan F. Hofmann, M.D., Telephone and fax: (858) 459-1754

Lee R. Hagey, Ph.D. Telephone (619) 543-2281

Abbreviations: ASBT, apical sodium co-dependent bile acid transporter; AUC, area under the curve; BSEP, bile salt export pump; CDCA, chenodeoxycholic acid; CMC, critical micellization concentration; CMT, critical micellization temperature; CMpH, critical micellization pH; DCA, deoxycholic acid; FATP, fatty acid transport protein; FGF, fibroblast growth factor; FGFR4, fibroblast growth factor receptor 4; FRET, Forster resonance energy transfer; FXR, farnesoid X-Receptor; GLP-1, glucagon-like peptide 1; MDR, multidrug resistance (protein); MRP4, multidrug resistance protein 4; MTBE, methyl tert-butyl ether; NTCP, Na+ taurocholate co-transporting polypeptide; OATP, organic anion transporting polypeptide; OST, organic solute transporter; PC, phosphatidylcholine; PFIC2, progressive familial intrahepatic cholestasis, type 2; SLCO, solute carrier organic anion; TGR5, transmembrane G protein-coupled receptor 4; UDCA, ursodeoxycholic acid.
Abstract

During the last 80 years there have been extraordinary advances in our knowledge of the chemistry and biology of bile acids. We present here a brief history of the major achievements as we perceive them. Bernal, a physicist, determined the x-ray structure of cholesterol crystals, and his data together with the vast chemical studies of Wieland and Windaus enabled the correct structure of the steroid nucleus to be deduced. Today, C_{24} and C_{27} bile acids together with C_{27} bile alcohols constitute most of the bile acid “family”. Patterns of bile acid hydroxylation and conjugation are summarized. Bile acid measurement encompasses the techniques of gas chromatography, HPLC, mass spectrometry, as well as enzymatic, bioluminescent, and competitive binding methods. The enterohepatic circulation of bile acids results from vectorial transport of bile acids by the ileal enterocyte and hepatocyte; the key transporters have been cloned. Bile acids are amphipathic, self-associate in solution and form mixed micelles with polar lipids -- phosphatidylcholine in bile, and fatty acids in intestinal content during triglyceride digestion. The rise and decline of dissolution of cholesterol gallstones by the ingestion of 3,7 dihydroxy bile acids is chronicled. Scientists from throughout the world have contributed to these achievements.

**Key words:** Bile acid transport, enterohepatic circulation, bile acid physical chemistry; bile acid analysis; bile acid metabolism
Table of contents

I. The ebb and flow of medical interest in bile acids 5

II. Bile acid structure: history and current concepts
   A. The carbon skeleton 13
   B. the A/B ring junction 16
   C. Hydroxylation sites 18
   D. Conjugation 25
   E. Structure-function relationships 28

III. Measurement
   A. Isolation and chromatographic separation 30
   B. Mass spectrometry 33
   C. Enzymatic and competitive binding techniques 34
   D. Diagnostic utility of plasma bile acids 35
   E. Determinants of bile acid profiles in body fluids 36

IV. Bile acid metabolism: transport and the enterohepatic circulation
   A. Bile acid transport 37
   B. Enterohepatic circulation in health 48
   C. Perturbations of the enterohepatic circulation 57
V. Physicochemical aspects of bile acids

A. Overview 58

B. Bile acids in aqueous solution 59

C. Solubilization of polar lipids by bile acid micelles 66

D. Physical chemistry of bile and cholesterol gallstone disease 70

E. Physical chemistry of fat digestion 73

VI. Dissolution of cholesterol gallstones

A. Gallstone dissolution by oral bile acids 75

B. Acceleration of dissolution by shock wave lithotripsy 79

C. Topical dissolution with organic solvents 80

D. Concurrent advances in gallstone management 81

VII. Epilogue: omissions and apologies 82
This article contains aspects of the history of bile acid research during the last eight decades. The authors, now in the twilight of their careers, are not trained historians, and there will be omissions, misconceptions, and provocative judgments. Let us hope the omissions are minor, misconceptions are rare, and provocative judgments stand the test of time. Let us also hope that this effort will be both useful and entertaining. Each of the topics discussed merits at least a full length article, so there will be of necessity in many instances consideration of only what we perceive to be the highlights.

I. The Ebb and Flow of Medical Interest in Bile Acids

After the elucidation of the true chemical structure of bile acids in 1932 (see below), there was little interest in bile acids in the Western world. One exception to this statement was the laboratory of Siegfried Thannhauser, who wrote the first textbook of metabolic biochemistry in Germany. He studied cholesterol and bile acid balance in the biliary fistula dog (1). During this time, bile acids were sold as liver tonics and laxatives, but there were no placebo-controlled studies showing efficacy. Indeed, bile acids were considered by the medical profession to have no useful therapeutic properties. The tri-oxo derivative of cholic acid (called “dehydrocholic acid”) was known to induce bile flow in animals (2), and was occasionally used to stimulate bile flow in patients; but again there were no controlled studies showing efficacy in hepatobiliary disease.

Albert Boehringer founded a pharmaceutical company in Ingelheim, Germany. His wife was a cousin of Heinrich Wieland, then Professor in Munich. Wieland was successful in persuading the company to begin the isolation and marketing of cholic acid from ox bile in 1917. Thus, cholic acid of high purity soon became available to the scientific community. During this time, Japanese students came to laboratories such as that of Wieland to pursue their doctoral studies. When they returned to Japan, they continued their studies on bile acid and bile alcohol structure. Their work, as well as that occurring in Europe is summarized in the book “Ueber die Chemie and Physiologie der Gallensäuren” authored by Shimizu in 1935 (3).
During the 1930’s and 1940’s, an enormous international effort went into defining the structure of the major hormonal steroids. When it was recognized that cortisone had a C-11 oxygen atom, it was logical to use deoxycholic acid (possessing a C-12 hydroxyl group) as a chemical precursor for the synthesis of corticosteroids. Deoxycholic acid (DCA) was easily isolated from bovine bile or synthesized from cholic acid, and it was not too difficult for a talented chemist to move the oxygen from C-12 to C-11. In 1946, H. Sarett (at the Merck Company) reported a complex synthesis (37 steps!) of cortisone from DCA (4), and this led to its commercial production on a small scale. Seven years later, Hench, a rheumatologist at the Mayo Clinic who worked closely with his colleague Kendall, an able steroid chemist, obtained a small supply of cortisone supplied by Merck (5). Hench showed that this compound caused a striking symptomatic improvement in patients with rheumatoid arthritis. The translational research of Hench and Kendall resulted in their being awarded the Nobel Prize in 1950 (6).

As a result of the truly exciting advent of cortisone in the treatment of rheumatoid arthritis, there was an enormous effort to develop a simple, efficient synthesis of this hormone from DCA. Soon, there was concern that the world’s supply of DCA (derived from cow and sheep bile) would be insufficient to meet the medical demand. Then, workers at Upjohn discovered that hydroxylation at C-11 could be achieved using a fungus. Other workers found a plant saponin that could be used as a substrate for the fungal hydroxylation at C-11, after which the side chain was easily altered to that of cortisone. (For details, see Fieser and Fieser) (7). DCA was no longer needed, and chemical interest in bile acids collapsed.

After the Second World War, a few laboratories pursued the search for new bile acids, as well as defining the metabolism of bile acids in mammals. The availability of $^{14}\text{C}$ and $^{3}\text{H}$ which could be incorporated into the bile acid molecule, together with the development of automatic liquid scintillation counters and chromatography enabled biotransformations to be identified and measured. Sune Bergström, then working in Lund, Sweden recruited a highly talented group of doctoral students -- Jan Sjövall, Henry Danielsson, Arne Norman, Sven Lindstedt, Bengt Samuelsson, Sven Eriksson, Bengt Borgström, among others -- who carried out fundamental studies of bile acid metabolism in a variety of
species including man (8). (It is said that Bergström chose his graduate students on the golf course. It is also said that the lights never went off at night in the Bergström laboratory).

Sjövall developed gas chromatography and then later, after he had moved to the Karolinska Institute in Stockholm, developed mass spectrometry for the measurement of bile acids (9). He used these techniques and others to define important aspects of bile acid metabolism over the next four decades (10). Norman developed a simple synthesis of conjugated bile acids, prepared the conjugates of all the major bile acids known at that time (11), and with Sjövall performed metabolic studies in animals which distinguished primary bile acids (made in the liver from cholesterol) from secondary bile acids (made from primary bile acids by intestinal bacteria). Their work proved that DCA was a secondary bile acid (12, 13). Lindstedt measured bile acid kinetics (pool size, turnover rate, and synthesis) in man (14). Danielsson identified key intermediates in bile acid biosynthesis (15). Samuelsson synthesized ursodeoxycholic acid (UDCA) and defined its metabolism (16). Eriksson showed that bile acid synthesis in the rat was under negative feedback control (17). Borgström with his colleagues Sjövall, Lundh, and Dahlquist, intubated healthy volunteers and defined the site at which individual nutrients are absorbed (18). This study proposed that the bile acid pool circulates about twice per meal. Borgström became director of the Department of Physiological Chemistry in Lund in 1958, replacing Bergström who had moved earlier that year to the Karolinska Institute in Stockholm with the bile acid group. Bergström and his pupil Samuelsson then left the bile acid field and turned their attention to prostaglandins. Their outstanding work in this field led to their receiving the Nobel Prize in 1982. However, at the Karolinska Hospital, Swedish scientists continued to work on bile acids, clinical and biochemical studies being conducted by Kurt Einarsson, Bo Angelin, Ingemar Björkhem, and Kjell Wikvall, among others.

In 1961, news of the work of Lack and Weiner (19) which showed unequivocally that the ileum was the site of conjugated bile acid absorption reached Sweden. Borgström directed the senior author, to perform studies showing that in man, conjugated bile acids are absorbed mainly in the ileum (20).
In London, Geoffrey Haslewood pursued the structure of bile acids in different species, based on his belief that bile acid composition was a powerful phenotype that would aid in the identification of evolutionary relationships (21). Edward Doisy, Chairman of the Department of Biochemistry at St. Louis University, having already won the Nobel prize for his work on vitamin K, pursued the identity of the 3,6,7-trihydroxy bile acids in rats and identified the $\alpha$- $\beta$- and $\omega$-muricholic acids (22). His student, William Elliott synthesized numerous bile acids and described their chemical and chromatographic properties (23).

In Japan, a school of bile acid studies was established by Tayei Shimizu (24), who returned to his native country after having spent three years in the laboratory of Heinrich Wieland. Japanese investigators pursued the structure of natural bile acids, the work continuing with Taro Kazuno and his pupil Takahiko Hoshita (25) and, in turn, with Hoshita’s student Mizuho Une (25, 26). Japanese workers also attempted to define the early steps in bile acid biosynthesis from cholesterol (27).

In the United States, James Carey, working at the University of Minnesota, was directed by his mentor, Cecil Watson, to pursue bile acid metabolism in man. He identified chenodeoxycholic acid (CDCA) as a major biliary bile acid (28) and proposed that lithocholic acid, its bacterial metabolite, caused liver injury in man (29). Edward (“Pete”) Ahrens, working with Lyman Craig at the Rockefeller University, showed that countercurrent distribution could be used to separate conjugated and unconjugated bile acids (30). Later, two of his postdoctoral fellows, Scott Grundy and Tatu Miettinen, developed a gas chromatographic method for measurement of fecal bile acids, whose output, in the steady state, is equivalent to bile acid synthesis (31). Grundy went on to a distinguished career in human nutrition (32). Miettinen continued to do sterol balance studies for much of his very productive career (33).

The first symposium devoted solely to bile acids was organized by Leon Schiff, a clinical hepatologist, who was one of the founders of the American Association for the Study of Liver Diseases. This symposium held in 1967 was quite exciting for its participants who are shown in Figure 1.
However, it is safe to say that the study of bile acids was pursued by only a small number of laboratories, some in Departments of Biochemistry and some in Departments of Medicine. Erwin Mosbach, one of the early workers in bile acid metabolism, once stated to his wife, “Whenever I go to the podium to give a paper on bile acids, everyone leaves the room”.

In 1965, the senior author, working in the laboratory of E.H. Ahrens, began feeding studies with cholic acid in a patient with severe hypercholesterolemia, and showed that cholic acid feeding was a potent suppressor of bile acid and cholesterol biosynthesis (34), based on measurement of fecal bile acids and sterols, using the newly developed gas chromatographic method for fecal bile acids that had been developed in this laboratory (32). It was logical to test CDCA, the other primary bile acid, but at that time, the world’s supply of pure CDCA was thought to be less than 10 grams, and the synthesis from cholic acid was difficult. However, in the 1960’s, a small English pharmaceutical company (Weddell Pharmaceuticals) began the manufacture of CDCA for unknown reasons. A kilogram was purchased for the senior author by the Mayo Clinic in 1967. Leslie Schoenfield returned to the Mayo Clinic in 1966 after having spent a year in the laboratory of Sjövall, and initiated a clinical trial with his fellow Johnson Thistle to test whether oral cholic acid or hyodeoxycholic acid would lower cholesterol in bile and ultimately induce cholesterol gallstone dissolution. The senior author persuaded Schoenfield to add CDCA to his protocol, and this study of Thistle and Schoenfield showed that CDCA feeding decreased biliary cholesterol saturation, whereas neither cholic acid nor hyodeoxycholic acid had any effect (35). In 1972, the first gallstone dissolution induced by the ingestion of CDCA was observed, initially at the Mayo Clinic (36), and later in London by a group led by Hermon Dowling (37). (This was not the first time that the efficacy of oral bile acids had been tested at the Mayo Clinic. In 1938, Philip Hench had fed a mixture of conjugated bile salts in an unsuccessful attempt to treat rheumatoid arthritis) (38).

The discovery that CDCA would induce gradual dissolution of cholesterol gallstones led to the next resurgence of interest in bile acids. For the very first time, CDCA was made in kilogram quantities by several manufacturers, and became the third bile acid available as a fine chemical.
In Japan, ursodeoxycholic acid (UDCA) had been brought to market in the 1950’s by Tokyo Tanabe as a liver tonic, based on the legendary therapeutic effects of bear bile. It was packaged as 10 mg tablets, and the recommended dose is unlikely to have had any pharmacodynamic effects. Two decades later in Japan, UDCA, in much larger doses, was observed to induce cholesterol gallstone dissolution (39, 40). UDCA soon replaced CDCA for gallstone dissolution throughout the world because UDCA showed an efficacy similar to that of CDCA, but in contrast, had virtually no hepatotoxicity. To provide UDCA for therapeutic trials performed around the world, UDCA was synthesized in kilogram quantities and became the fourth bile acid to become available as a fine chemical. A Japanese perspective on the discovery of UDCA as a therapeutic agent is available (41).

As discussed later, dissolution of cholesterol gallstones by ingestion of UDCA began to decline in the 1990’s because of the development of laparoscopic cholecystectomy which promised a rapid cure for the problems of gallstone disease, irrespective of gallstone type. Once again, bile acids seemed to have little medical value, and interest in their biology and chemistry declined.

Nonetheless, the ready availability of UDCA meant that it continued to be used for gallstone patients who were poor operative candidates. Ulrich Leuschner, a hepatologist based in Frankfurt, Germany, noted that UDCA ingestion improved biochemical parameters in patients with primary biliary cirrhosis (42). UDCA was subsequently shown to slow disease progression in primary biliary cirrhosis, initially by Raoul Poupon in Paris (43), and later by Keith Lindor and his colleagues at the Mayo Clinic (44). UDCA treatment of primary biliary cirrhosis is now standard of care. UDCA was also shown to be effective in cholestasis of pregnancy (45). UDCA has been widely used for other cholestatic diseases and to restore bile flow following liver transplantation, but proof of efficacy in these conditions based on placebo-controlled trials has been uncommon.

At present, it is estimated that about 1000 metric tons or about 1 million kilograms -- of cholic acid are produced globally from bovine bile, the majority of which is used for the production of UDCA (S. Yorke,
New Zealand Pharmaceuticals, Ltd.; personal communication). Bile acids are also used as precursors for the synthesis of nuclear receptor agonists such as obeticholic acid (the 6α-ethyl derivative of CDCA), as building blocks for other molecules (46), and as key constituents of some bacteria media. Bile acids have even used in the electronic industry as precursors of photoresist molecules in integrated circuits (47). (This unlikely technology transfer from a biological laboratory to the electronic world resulted from one of the early chemists in our laboratory, Rick DiPietro, taking a position in the electronics industry and envisioning a truly novel use for bile acids).

In 1984, the results of the Coronary Primary Prevention Trial showed that the ingestion of cholestyramine, the first bile acid sequestrant to be marketed, lowered plasma cholesterol levels and reduced cardiovascular events. This was of considerable interest to the cardiovascular community, as the study provided the first convincing evidence for the validity of the cholesterol hypothesis (48). Nonetheless, the demonstrated efficacy of treating hypercholesterolemia with a bile acid sequestrant had a rather modest effect on clinical practice because of the poor tolerability of cholestyramine and the emergence of the statins which were still more potent and usually well tolerated.

The current rebirth of interest in bile acids arises from the astonishing discovery of their signaling properties. Bile acids were found to be the ligand for the orphan nuclear receptor FXR (farnesoid X-receptor) (49, 50, 51), meaning that bile acids regulate the activity of many genes (52, 53, 54). Bile acid derivatives that are potent FXR agonists have been shown to induce the synthesis and basolateral release of the protein fibroblast growth factor 15/19 (FGF19) from the ileal enterocyte (55). FGF19 travels in portal venous blood to the liver and down-regulates bile acid biosynthesis (55, 56). FGF19 promotes hepatocyte growth (57), and FXR agonism also has antidiabetic effects mediated by multiple mechanisms (58), as well as anti-inflammatory activities (59). Several potent FXR agonists have been synthesized (53, 60) and one of these, obeticholic acid (61), is now being tested for its efficacy in primary biliary cirrhosis and non-alcoholic steatohepatitis (62).
Two Japanese groups discovered a transmembrane G-coupled protein receptor (TGR5) activated by bile acids (63, 64). TGR5 has a wide tissue distribution, being most heavily expressed in the enteric nervous system (65); it is also present in spinal cord neurons (66) and the primary cilia of cholangiocytes (67). The consequences of this activation include release of the hormone glucagon-like peptide 1 (GLP-1), which has appetite-suppressing and antidiabetic effects (68). TGR5 also has effects on hepatic blood flow (69), distal intestinal motility (70), and in addition, has anti-inflammatory effects (71) that are under active investigation.

Discovery of the signaling properties of bile acids is certainly a new paradigm in the bile acid field. Bile acids have become hormones. Now highly talented molecular biologists in both academe and the biotechnology industry are pursuing bile acid biology. Bile acid signaling has been discussed only briefly in this review, as its aim is to focus on events in the past.

As bile acids have been the subject of scientific study for more than 150 years, the bile acid literature is vast. The early history of bile acid discovery is available in the two monographs by Harry Sobotka (72, 73), as well as the masterly text on steroids by the Fiesers published in 1955 (7). A small pharmaceutical company, Dr. Falk Pharma, based in Freiburg, Germany was the first European company to bring CDCA to market for gallstone dissolution. Since 1972, the Falk Foundation, funded by Dr. Falk Pharma, has hosted biennial meetings on bile acid biology, chemistry, and therapeutics. The proceedings of these meetings, usually published in book form, serve to document the advances that have occurred in the last four decades. G.A.D. Haslewood’s slim monograph summarizes current knowledge in 1966 (74) and Haslewood later authored a monograph on the biological importance of bile salts (75). The Elsevier Encyclopedia of Organic Chemistry (76) contains detailed information on the chemistry of bile acids and their derivatives. A four volume series “The Bile Acids” edited by Nair and Kritchevsky (77) treats many of the topics discussed in this article in much greater detail. A book entitled “Sterols and Bile Acids” is Volume 12 of the series New Comprehensive Biochemistry. This volume, published in 1985, has 6 chapters dealing with varying aspects of bile acids and bile alcohols (78). A multiauthor book describing
many aspects of bile acid research with particular emphasis on Italian contributions was published in 2000 (79). The senior author (80) and Hermon Dowling (81) have both written overviews of the field from a medical standpoint in 1985. The Journal of Lipid Research published a number of excellent reviews on the biology and chemistry of bile acids a few years ago.

The remainder of this historical review will discuss selected topics in more detail with an emphasis on structure-activity relationships. The review will not summarize our extensive work on bile acids in different vertebrates, as our findings have been published elsewhere (82). Other topics that have been omitted are discussed in the epilogue.

II. Bile acid structure: History and current concepts

A. The carbon skeleton

1. The steroid nucleus

In December, 1929, a unique event occurred in Stockholm. Two Nobel prizes in Chemistry were awarded. In 1927, no prize in Chemistry had been awarded because the Nobel Prize Selection Committee could not agree on a suitable candidate. During the following year, the committee decided to award the 1927 prize to Heinrich Wieland and the 1928 prize to Adolf Windaus. In his Nobel lecture, Wieland reviewed his laborious attempts to define the structure of bile acids and presented his current vision of the structure of cholic acid (83). Windaus presented his proposed structure of cholesterol as well as his brilliant studies on the formation of vitamin D2 from cholesterol by radiation (84). These two extraordinarily talented chemists, using the most primitive of chemical techniques, had slowly developed the idea that both cholesterol and bile acids share a four ring structure with a side chain of five carbon atoms for bile acids and eight carbons for cholesterol. However, the structures of the steroid nucleus shown in their Nobel lectures were not correct, and each of these giants described their structures as “probable”.
The path to the correct structure came from an unlikely origin. The polymath Desmond Bernal, working in the Mineralogical Museum in Cambridge, England decided together with his colleague William Astbury to undertake X-ray diffraction studies of organic molecules. Bernal, nominally a physicist, began his studies in 1931. He studied ergosterol, cholesterol, and vitamin D and for all three of these compounds obtained a probable structure of a flat molecule. He published his findings as a note in Nature in 1932 (85). This report caught the attention of Rosenheim and King in London, and Wieland and Dane in Munich. Later, in that same year, Wieland and Dane followed by Rosenheim and King published the application of Bernal’s pioneering studies to the extensive literature on bile acid structure (86). Each group proposed the cyclopentanoperhydrophenanthrene structure for the steroid nucleus. Their final structure has stood the test of time with confirmation by both X-ray diffraction (87) and a total synthesis of CDCA (88). Figure 2 shows the structure of CDCA and the numbering system for the steroid nucleus and side chain that was proposed in the 1930’s and has been used ever since.

Bernal’s X-ray diffraction studies were the first biological application of this revolutionary technique. Bernal later used X-ray diffraction to deduce the structure of other organic molecules, and his pupils such as Dorothy Crowfoot Hodgkin extended the work, its ultimate triumph being the helical structure of DNA and RNA by Watson and Crick. Bernal’s left wing political views and his Bohemian life style probably kept him from being awarded a Nobel Prize (89,90).

2. The side chain

The structure of the C₅ side chain of the common natural C₂₄ bile acids – isopentanoic acid or more specifically γ-methylbutanoic acid – was elucidated by Wieland, and like the structure of the steroid nucleus, has also been confirmed by X-ray diffraction (87) and total synthesis (88).

The structure of the C₈ side chain in C₂₇ bile alcohols and bile acids was more problematic. It was not clear initially that these molecules possessed the side chain of cholesterol because it was quite uncertain whether these compounds and even the common natural C₂₄ bile acids were derived from
cholesterol. The uncertainty arose from the failure of a dietary cholesterol load to increase bile acid output in the biliary fistula dog. George Whipple, a leading physiologist, wrote in his 1922 review on the constituents of bile (91). “Cholesterol....has often been suspected as the precursor of cholic acid, ... but careful experiments negative [sic] this interesting suggestion”. Nonetheless, Shimizu in his book “Ueber die Chemie und Physiologie der Gallensäuren”, published in 1935 (3) noted that, “Interesting intermediates in the biological catabolism of cholesterol to bile acids can be found in scymnol and tetraoxybufostan [both C27 bile alcohols]”. Even earlier, in 1923, Ruzicka had proposed that bile acids are formed from cholesterol (92).

Proof that cholesterol was converted to bile acids had to await the introduction of isotopes into biochemistry by Rudolph Schoenheimer (93). Bloch, Berg, and Rittenberg prepared cholesterol labeled with deuterium, gave it to a dog with a surgical connection between the biliary tract and the renal pelvis; in this chronic preparation bile is excreted in the urine. Cholic acid was isolated from urine and shown to have incorporated deuterium into its structure, thus establishing the conversion of cholesterol to C24 bile acids (94). Figure 3 shows the structure of cholesterol and the changes that occur as it is converted to CDCA. (Although Konrad Bloch received the Nobel Prize in 1964 for his extraordinary work on fatty acid and sterol synthesis, his Nobel lecture does not even mention this key experiment).

The first bile acid with a longer side chain than the common natural C24 bile acids was isolated from the bile of the toad *Bufo vulgaris formosus* by Shimizu and Oda in 1934 (95). The structure of this bile acid was established by Takahiko Hoshita *et al* as a trihydroxy bile acid with a double bond at C-22 and a carboxyl group at C-24 (25). Thus this bile acid is a C28 bile acid having one more carbon than its precursor cholesterol. Other C28 bile acids with a methyl group on the side chain have been reported (26), but are quite uncommon in nature.

C27 bile acids are much more common (82). They were isolated from the bile of a number of early evolving vertebrates. In particular, bile from the alligator, which was readily available in large
amounts (studies at that time began with up to 5 liters of bile), became the object of intense study as it contains a great variety of \(C_{27}\) bile acids (96, 97).

The carbon skeleton of the side-chain of \(C_{27}\) bile acids was assumed to have remained unchanged from that of the parent compound, cholesterol, but the location of the carboxyl group was unclear. In 1952, Bridgwater and Haslewood synthesized a defined side-chain starting from the steroid nucleus. They then showed that their product was identical to \(3\alpha,7\alpha,12\alpha\)-trihydroxy-\(5\beta\)-cholestan-27-oic acid, the major bile acid isolated from the bile of crocodiles and alligators, and they confirmed its structure by X-ray diffraction (98). Subsequently, the laboratory of Takashi Iida reported the synthesis of the \(25R\) and \(25S\) stereoisomers of this compound (99).

The first \(C_{27}\) bile alcohol to be isolated was scymnol, obtained from the bile of the Greenland shark (\textit{Scymnus borealis}) by Olof Hammarsten, working in Sweden in 1898 (100). The finding that the molecule had 27 carbon atoms is a tribute to the precision and accuracy of the combustion techniques that were used at that time to determine elemental analysis. A half century later, another \(C_{27}\) bile alcohol was isolated from the bile of the American bullfrog (\textit{Rana catesbeiana}) by Kazuno et al (101). Subsequently, Haslewood (102) found the same bile alcohol in the European common frog (\textit{Rana temporia}), and gave it the common name “ranol”. Our extensive survey of vertebrate bile acids (82) has established that most bile alcohols have an \(\alpha,\epsilon\)-dimethyl-hexanoic acid side chain. The exceptions identified to date are \(C_{26}\) and \(C_{25}\) alcohols that are present in the bile of some frog species (25, 26), as well as a \(C_{24}\) bile alcohol, petromyzonol that occurs in the lamprey as a disulfate (103,104). Trace proportions of \(C_{24}\) bile alcohols with the nuclear structure of cholic acid and CDCA are found in the bile of many species (105).

**B. The A/B ring junction**

In most natural bile acids whether \(C_{24}\) or \(C_{27}\), the junction of the A and B rings is \textit{cis}. This steric conformation is indicated by depicting the bond connecting the C-5 hydrogen atom to the A/B ring.
juncture as a solid line or wedge. The much rarer "allo" bile acids (A/B ring junction is trans) are usually denoted by a dashed line coming from the ring junction to the hydrogen atom. For the non-steroid chemist, it is not easy to grasp that the orientation of the C-5 hydrogen atom indicates the type of A/B ring junction, and thus determines the shape of the two markedly different bile acid epimers. The structures of cis and trans decalin, model compounds for the A and B rings of bile acids are shown in Figure 4. Every bile acid and bile alcohol has, in principle, a 5α and a 5β epimer.

Both Wieland and Windaus were fully aware of the problem of bile acid geometric isomerism at the time of their Nobel Prize lectures. A detailed review of some of the experimental work elucidating the nature of the A/B ring junction in various bile acids is available in Fieser and Fieser’s comprehensive text on steroids (7). Probably, it was Windaus who introduced the prefix allo to denote 5α (A/B trans) bile acids.

Elliott has summarized the chemical and biological properties of allo bile acids in his chapter in The Bile Acids, volume I (106). One way to distinguish 5α bile acids from 5β bile acids is to prepare their per-oxo derivatives, which are readily distinguished using optical rotatory dispersion (107).

C24 allo bile acids are the dominant primary bile acids only in lizards, although they are present in trace proportions in the bile of many species (82). If cholestanol (the saturated 5α-derivative of cholesterol) is administered to rabbits, it is converted to 5α-bile acids (107). Allocholic acid, the biological precursor of allodeoxycholic acid, has been prepared synthetically and some its biological properties explored (108). Allo bile acids have also been identified in fecal bile acids (109), presumably generated by bacterial enzymes acting on the 3-oxo-Δ4,6 intermediate that is formed during the process of C-7 dehydroxylation (110). Enantiomeric bile acids, in which the steric configuration of all ring junctions is reversed, have recently been synthesized (111).
Bile alcohols (C27) with a 5α configuration are not uncommon in nature. The simplest biotransformation of cholesterol (Δ5) into a bile alcohol results in a 5α nucleus, and many early evolving vertebrates form A/B trans bile alcohols (25, 26, 82).

The preceding discussion thus indicates that there are only four carbon skeletons in most natural bile acids and alcohols – the C27 cholestan skeleton (with either A/B cis or A/B trans structure) and the C24 cholane skeleton (with either A/B cis or A/B trans A/B structure). The side chain of each of the four skeletons may end in either a primary alcohol group or a carboxyl group. As noted above, C24 bile alcohols are very rare in nature. Thus, to simplify bile acid and alcohol nomenclature, it is useful to divide all “cholanoids” and “cholestanoids” into only three great classes – C27 bile alcohols, C27 bile acids and C24 bile acids. In each of these classes, there are 5α and 5β epimers. Within these three great classes, and for a given A/B ring junction, the pattern of hydroxylation – on the steroid nucleus and side chain or both -- defines each individual bile alcohol or bile acid.

C. Hydroxylation sites

1. The steroid nucleus

   a. C24 primary bile acids

   Bile acid nomenclature, i.e. the trivial names of bile acids, is awkward because it is based on cholic acid. Ox bile from the slaughter house was readily available to chemists, and Strecker is given credit for isolating cholic acid and obtaining the correct empirical formula of C24H40O5 by combustion analysis in 1848 (112). Mylius isolated DCA in 1886 and the term “deoxy” was attached as a prefix because DCA had one less oxygen atom by combustion analysis (113). When CDCA was isolated from goose bile, it was named Chenosäure (cheno acid) for its animal source of origin (cheno denotes goose in Greek) (114), and some years later the adjective “deoxy” was added, based on its elemental composition, giving rise to the polysyllabic name of chenodeoxycholic (acid). Other trihydroxy bile acids were named for their animal source: muricholic acid (from rodents), and hyocholic acid (from pigs).
We have proposed (82) that the root C24 bile acid should be CDCA, as every bile acid must have a hydroxyl group at C-3 (from the C-3 hydroxyl group of cholesterol) and a hydroxyl group at C-7, as 7α-hydroxylation of cholesterol is the rate limiting step in the major pathway of bile acid biosynthesis. Thus, all natural primary C24 trihydroxy bile acids can be considered as derivatives of CDCA. The only exception to this useful convention occurs in those few mammals (nutria, beavers, bears, and some cimorphs) where UDCA, the 7β-hydroxy epimer of CDCA, is a primary bile acid. The route of formation of this 7β-hydroxy epimer of CDCA is unknown. One can speculate on the existence of a cholesterol 7β-hydroxylase or a pathway that results in the epimerization of the C-7 hydroxy group.

The most frequent site (in terms of the number of vertebrate species) of additional hydroxylation of CDCA is at C-12 in cholic acid. The second most common site of hydroxylation is probably at C-16, because so many avian species have the 3α,7α,12α,16α-trihydroxy bile acid as a major biliary bile acid (82). Another common site of hydroxylation is at C-6, occurring in hyocholic acid (pigs) and the three muricholic acids (rats and mice). Other “third site” (nuclear) hydroxylations occurring in biliary bile acids of the healthy animal are at C-1 (α or β), C-2 (β), C-4 ( ), C-5 (β), C-9 (β), C-11 (α), C-15 (α), and C-19 (β). Table 1 summarizes the species in which these rare biliary bile acids occur, and the appropriate citation.

C24 bile acids with four hydroxyl groups on the steroid nucleus occur rarely. Some snakes have 3α,7α,12α,16α-tetrahydroxy bile acids (82). Tetrahydroxy bile acids of unknown structure are found in trace proportions in the bile of many vertebrate species. After experimental ligation of the common bile duct, tetrahydroxy- and even penta-hydroxy C24 bile acids occur in the plasma and urine of mice (115). Such bile acids also occur in mice when the canalicular bile salt export pump has been deleted genetically (116).

An ad hoc committee of bile acid workers met in Freiburg in 1980 under the auspices of the Falk Foundation. The group proposed working rules for bile acid nomenclature and bile acid abbreviations
that were published and noted by the IUPAC nomenclature committee. At the suggestion of Gerhart Kurz of the University of Freiburg, it was agreed that there would be no more trivial names, except for conversational purposes (117).

b. C₂₄ secondary bile acids

As noted before, bile acids are modified by bacterial enzymes in the distal intestine. Modification of the hydroxyl groups forms new bile acids that are termed “secondary”. This distinction has limitations, because bile acids such as UDCA or certain of the hydroxy-oxo bile acids can be both primary and secondary. A consideration of secondary bile acids is important because they can be major biliary bile acids. After their formation in the distal intestine, they are absorbed in part. They return to the liver in portal venous blood, and traverse the hepatocyte where they may or may not undergo further metabolism before joining the pool of primary bile acids (formed de novo from cholesterol) in the enterohepatic circulation.

The major bacterial biotransformations of conjugated bile acids are deconjugation, 7-dehydroxylation, oxido-reduction, epimerization, and side chain desaturation (detailed in 110). Deconjugation, i.e. hydrolysis of the amide bond between the terminal carboxylic acid group of the bile acid and the amino group of taurine or glycine, is mediated by many bacterial species and does not require anaerobic conditions. Unconjugated bile acids formed by bacterial deconjugation are still considered primary bile acids.

In the large intestine, anaerobic modifications of the hydroxyl groups occur. The most striking change is 7-dehydroxylation that occurs via a 3-oxo-Δ⁴,6 nucleotide intermediate (110). The 7-dehydroxylation product of cholic acid is DCA; that of CDCA or UDCA is lithocholic acid. The 7-deoxy derivative of hyocholic acid and ω-muricholic acid is termed hyodeoxycholic acid and that of α- and β-muricholic acid is termed murideoxycholic acid. To date, the C-7 deoxy derivative of “avicholic” (3α,7α,16α-trihydroxy) has not been identified, but bacterial 16-dehydroxylation has been reported for some steroids (118), and is
likely to occur in some birds. Such 16α-dehydroxylation would result in the formation of CDCA, that in turn would likely undergo 7α-dehydroxylation to form lithocholic acid.

During transit through the large intestine, any hydroxyl group can be oxidized and then reduced. If the reduction is to the other possible epimer, a new epimer will be formed. 3β-Hydroxy bile acids (“iso” bile acids) are major fecal bile acids in man (119). 7β-Hydroxy bile acids and even 12β-hydroxy bile acids are also formed, but to a far lesser extent. Elimination of the 3-hydroxy group with formation of a Δ2 or Δ3 bile acid may also occur (120).

Secondary bile acids may undergo additional hydroxylation during hepatocyte transport. Such rehydroxylation at C-7 generates the original primary bile acid that had undergone bacterial 7-dehydroxylation in the intestine. This rehydroxylation does not occur in man and the cow, occurs to a limited extent in the rat and hamster, and is complete in the prairie dog (121). Alternatively, secondary bile acids may undergo hydroxylation at a new site, generating a novel “tertiary” bile acid. One example is 16α-hydroxylation of DCA to form “pythocholic” acid (3α,12α,16α-trihydroxy) that occurs in the python (75). In the wombat, lithocholic acid undergoes hydroxylation at C-15 (122). “Iso” (3β-hydroxy) bile acids are epimerized to 3α-hydroxy bile acids during transport through the hepatocyte (123,124). Biotransformation of the 3β-epimers of the muricholic and hyocholic acid families has not been reported, but is likely to occur. In man, the 7β-hydroxy group of UDCA does not undergo epimerization during transport through the hepatocyte, and the small proportion of UDCA in the biliary bile acids is likely to be formed from CDCA in the distal intestine by bacteria. Bacterial modification of bile acids and subsequent hepatic biotransformation has been termed by us, “damage and repair”.

The percent of secondary bile acids in biliary bile acids is usually less than 50%, but there are exceptions. The most remarkable example is in the rabbit, in which DCA constitutes 80% of the circulating bile acids (82). In some snakes, the 3α,12α,16α-trihydroxy bile acid, which can be formed by 16α-hydroxylation of DCA, exceeds 50% (82).
c. C_{27} bile acids

Nuclear hydroxylation in C_{27} bile acids occurs at C-3 and C-7 (the default structure), and also at C-12 in most species. In the Andean condor, hydroxylation at C-16 occurs (82). We have identified a 1\beta,3\alpha,7\alpha-trihydroxy C_{27} bile acid in the bile of the tinamou, an ancient flightless bird (125). A tetrahydroxy bile acid with hydroxyl groups at C-3, C-7, and C-12 plus a 2\beta-hydroxy group has been isolated from the bile of the large freshwater fish *Arapaima gigas* (126). It seems likely that there are additional sites of nuclear hydroxylation in C_{27} bile acids yet to be discovered.

C_{27} bile acids undergo 7-dehydroxylation in the alligator (97), but the extent of 7-dehydroxylation of C_{27} bile acids in other species has received little attention.

d. C_{27} bile alcohols

In C_{27} bile alcohols, default nuclear hydroxylation occurs at C-3 and C-7; additional hydroxylation occurs at C-12, and rarely at C-6 or C-16 (25,26, 127). The C-3 hydroxyl group has a \beta-configuration in latimerol and myxinol, meaning that it remains unchanged from that of cholesterol. Hydroxylation at C-16 occurs in myxinol. The bile of *Arapaima gigas* contains a bile alcohol with a \beta-hydroxy group at C-2 in addition to hydroxyl groups at C-3, C-7, and C-12, so that C_{27} bile alcohols with four hydroxyl groups on the nucleus do occur (126).

No C_{27} bile alcohols are recognized as “secondary”, i.e. with hydroxyl groups modified by bacterial enzymes. However, the presence of 7-deoxyscymnol in the Megamouth shark, as reported by Ishida et al (128), suggests that bacterial dehydroxylation at C-7 may also occur for C_{27} bile alcohols.

2. The side chain

a. C_{24} bile acids
With only four available carbons in the side chain, C_{24} bile acids are likely to undergo side chain hydroxylation only at C-23 or C-22, each of which has potential R and S epimers. The most common C-23 hydroxy bile acid is termed “phocaeholic acid” and has the nucleus of CDCA. (The name “phoca” is given to the marine mammal family from which Hammarsten (129) isolated the bile acid). The structure of phocaeholic acid was later established as the 23-hydroxy derivative of CDCA by Windaus and van Schoor (130). The R and S epimers have been synthesized (131,132) and the R form identified as the naturally occurring epimer (131). In addition to marine mammals, phocaeholic acid also occurs in wading birds and some song birds (82). The 12-hydroxy derivative of phocaeholic acid also occurs in marine mammals and is therefore a natural tetrahydroxy- bile acid.

A 3α,7α-dihydroxy bile acid with a hydroxy group at C-22 was shown to occur in the fish *Parapristipoma trilineatum* by the Hoshita laboratory (133) and given the awkward name of haemulcholic acid. The compound was synthesized by both the Hoshita laboratory (134) and the Pellicciari laboratory (135), and the S epimer was shown to occur in nature. Some of its physicochemical and biological properties were characterized by Roda et al (135). Haemulcholic acid was also identified in freshwater fish by Haslewood during his Zaire River expedition (75).

Side chain desaturation forming Δ^{22} bile acids is a major pathway in rats (136,137) and cavimorphs (82). Such bile acids could arise from further β-oxidation of the side chain in the liver and/or by bacterial action in the distal intestine.

b. C_{27} bile acids

Elucidation of the structure of the C_{27} bile acids with hydroxyl group substituents on the C₈ side chain was not simple, as side products such as lactones may be generated during alkaline hydrolysis. Unfortunately, an enzyme preparation that hydrolyzes the amide bond of C_{27} taurine conjugated bile acids is not available at present (In contrast, cholylglycyl hydrolase, which rapidly cleaves the amide bond in C_{24} conjugated bile acids, is available commercially).
So far, all reported C\textsubscript{27} bile acids have only a single hydroxyl group in the side chain. It occurs at C-22 in turtles, at C-24 in lizards, and at C-26 in bullfrogs (25, 26); each of these bile acids has four potential epimers. The four cholestanolic acid epimers with a hydroxyl group at C-24 have each been synthesized by Une et al. (138). The most common diastereoisomer is varanic acid, which has the 24\texttextsubscript{R}, 25\texttextsubscript{R} configuration. In collaboration with the Iida and Une groups, we have found that the 24\texttextsubscript{R}, 25\texttextsubscript{S} epimer is present in the bile of several varanids (139).

c. C\textsubscript{27} bile alcohols

As noted earlier, Hammarsten determined that he had isolated a C\textsubscript{27} bile alcohol from the bile of the Greenland shark (100). The challenge of determining the structure of this compound would occupy chemists for many decades. Windaus \textit{et al}. (140) initially took up the problem of determining the structure of Hammarsten’s compound (named scymnol) and proposed that it contained one primary and three secondary hydroxyl groups, with an epoxy ring. The three secondary hydroxyl groups were found to be located in the sterol nucleus (at positions 3\text{\textalpha}, 7\text{\textalpha}, and 12\text{\textalpha}) as scymnol could be converted into cholic acid by partial oxidation of the side-chain (141,142). Since the side-chain of cholesterol has a C-27-methyl group, different paired combinations of carbons 23, 24, and 25 were proposed to contain the epoxy group, with the primary alcohol group being on C-27. Bridgwater and Haslewood showed (using Hammarsten’s original sample) that the epoxy group was an artifact of alkaline hydrolysis, and that the side-chain of scymnol actually contained two secondary hydroxyl groups and a primary hydroxyl group bonded to sulfate (143,144). In 1961, Cross used \texttextsuperscript{1}H-NMR to elucidate the structure of both scymnol and anhydroscymnol, the epoxide formed during alkaline hydrolysis (145). He found that the side chain hydroxyl group of scymnol was unexpectedly on C-26 (the only carbon with a single proton), consistent with a side-chain structure having hydroxyl groups at C-24 and C-26, and a primary hydroxyl group at C-27. The structure of scymnol was ultimately established as 3\text{\textalpha}, 7\text{\textalpha}, 12\text{\textalpha}, 24, 26, 27-hexahydroxy-5\text{\textbeta}-cholestan by comparison with a synthetic standard (143). Carbon-24 is chiral, and the configuration of the 24-hydroxyl group was determined to be 24\texttextsubscript{R}, in the crystallographic study of Ishida et.al. (146). This
group has also found that whereas most shark species have the (24R,25R)-epimer of scymnol, a few species also form the (24R,25S)-epimer (147). Other named C_{27} bile alcohols are tabulated in our previous review (82). Additional information is available in the cogent review of Une and Hoshita (26).

Bile alcohols, just as bile acids, have been named for the species in which they occur, and the name gives no clue whatsoever to the structure of the bile alcohol. It is convenient to think of the trihydroxy bile alcohol with α-hydroxyl groups at C-3, C-7, and a primary alcohol function at C-27 as the “root” C_{27} bile alcohol, and to look on other bile alcohols as derivatives of that simplest bile alcohol. At present, this “root” bile alcohol has no trivial name, but has been synthesized by the Mosbach group (148).

D. Conjugation

1. N-acyl amidation with taurine or glycine.

In virtually all species, bile acids, whether C_{24} or C_{27}, are found in bile in N-acyl amidated form, i.e. conjugated in amide linkage with taurine (or a taurine derivative) or far less commonly with glycine. Taurine conjugation is the rule in fish, amphibians, reptiles, and most avian species. It also occurs to a variable extent in mammals (82). In some angelfish species, N-methyltaurine conjugates are present (149), the N-methyltaurine thought to be of dietary origin. Similarly, in some marine fish, bile acids are conjugated with cysteinolic acid, a hydroxy-ethyl derivative of taurine that is of dietary origin (150). Glycine conjugation occurs in a few avian species, some bovids, caviomorphs, and primates, including humans (82). Sarcosine (N-methylglycine) conjugates have been synthesized (151,152,153) and shown to be more resistant to bacterial deconjugation during enterohepatic cycling (154). Some workers have considered bile acid amidation to be the final step in bile acid biosynthesis (155). It seems more logical to us to define the bile acid biosynthesis pathway as ending with a mature bile acid or bile alcohol before it undergoes a phase II biotransformation step such as N-acylamidation, sulfation, or glucuronidation.

The process of conjugation occurs both in hepatocytes and in cholangiocytes (156). However, conjugation in the hepatocyte greatly exceeds that occurring in cholangiocytes. The conjugation process
begins with the formation of a bile acid adenylate (157) followed by formation of the coenzyme A thio-
ester. A bile acid aminotransferase then transfers the CoA bile acid to the amino group of glycine or
taurine; the human enzyme has been cloned and shown to mediate conjugation with both glycine and
taurine (158). Conjugation of newly synthesized bile acids occurs in peroxisomes (159), and recent work
suggests that “reconjugation” of unconjugated bile acids returning from the intestine also occurs in these
organelles (160). Other than their pKₐ value, which in turn is responsible for the pH-solubility profile (see
later), there is little difference in the biological properties of glycine and taurine conjugated bile acids,
except that taurine conjugates are deconjugated by bacteria more slowly than glycine conjugates
(161,162).

That bile acids are present in bile in “conjugated” form was recognized in the 19th century. At that
time, neither the structure of the bile acid nor the site of linkage was known. This resulted in bile acid
conjugates having a misleading name, in that “taurocholate” implies that the cholic acid moiety is capable
of ionization. We and others have suggested that the name “cholytaurine” (analogous to glycylglycine)
would be more appropriate (117), but few investigators have followed this proposal. The extremely low
proportion of unconjugated bile acids in biliary bile acids has led some investigators to conclude that the
conjugation process is highly efficient. However, unconjugated bile acids, especially dihydroxy bile
acids, if secreted into bile are rapidly absorbed by the biliary ductules (163,164), resulting in their
removal from ductal bile. Thus, the state of conjugation in canalicular bile can be considerably less than
that of ductal bile.

2. Other types of bile acid conjugation

When the term “conjugated” was applied to bile acids, it denoted N-acyl amidation with taurine or
glycine. When bile acid sulfation and glucuronidation were discovered to be important pathways under
some circumstances, it was necessary to distinguish different modes of “conjugation” for bile acids. In
principle, there are six modes of bile acid conjugation: N-acyl amidation (using glycine, taurine, or
possibly other amino acids), sulfation, glucuronidation, conjugation with glutathione, and conjugation with N-acetylglucosamine or glucose, and bile acids may have more than one mode of conjugation. In vertebrates, the only bile acid sulfates that occur in appreciable proportions in biliary bile acids are the 3-sulfate of lithocholyglycine and lithocholyltaurine amidates that occur in man (165) and probably great apes (166). In patients with severe cholestasis, bile acid sulfation (mostly at C-3) followed by urinary excretion becomes a major pathway (167,168). In the bile duct ligated hamster, esterification with sulfate occurs at C-7 (169).

In the Mayo Clinic studies of Thistle and Schoenfield, in which bile acids were administered to gallstone patients (35), one group received hyodeoxycholic acid (3α,6α-dihydroxy-). Since hyodeoxycholic acid should be absorbed efficiently from the small intestine by passive mechanisms, its absence in biliary bile acids was completely mystifying. In 1978, B. Almè and J. Sjövall reported a patient with malabsorption who excreted in urine 6-hydroxy bile acids that were present as O-linked (ethereal) glucuronides. (170). Five years later, Sacquet et al, working in the laboratory of Reca Infante in Paris, reported that when hyodeoxycholic acid was administered to human subjects, it was largely excreted in urine as its glucuronide (171), explaining its absence in biliary bile acids. This group did not know the site of glucuronidation but proposed that it was at C-3.

Four years later, Anna Radominska-Pyrek and colleagues showed that human microsomes rapidly glucuronidate hyodeoxycholic acid at C-6 (172). Thus the curious fate of hyodeoxycholic acid in humans was finally settled. It presumably results from defective N-acylamidation, and hyodeoxycholic acid is consequently metabolized in the same manner as a C_{24-nor} (C_{23}) dihydroxy bile acid (see below), i.e. by glucuronidation in the endoplasmic reticulum. To date, this pathway has not been identified in any other animal. When hyodeoxycholic acid is given to rats, it is probably conjugated with taurine and not glucuronidated (173).
When C23 dihydroxy-bile acids, are administered to animals including man, little amidation occurs, as both the CoA ligase and the glycine/taurine amino acid transferase act to only a very limited extent on bile acids with a shortened (C4) side-chain (174). Such bile acids undergo glucuronidation both at C-3 and C-23 (163,164). In man, when norUDCA is given, it forms the C-23 ester glucuronide which is excreted in both bile and urine (175). In rats, N-acyl amidation is also much less efficient for α-hydroxy bile acids (131,132) and hydrophilic epimers of cholic acid, for example 3α,7β,12β-trihydroxy-cholanoate (176).

In patients ingesting UDCA, UDCA also undergoes conjugation with N-acetylglucosamine at C-7, in addition to amidation with glycine or taurine at C-24 (177). In Niemann-Pick disease type C, a novel Δ5 3β,7β-dihydroxy cholestanolic acid is formed (178); this compound may be considered an acidic 7β-hydroxy derivative of cholesterol. This compound undergoes sulfation at C-3, conjugation with N-acetylglucosamine at C-7 (178), and amidation with glycine or taurine at C-24. In both circumstances, the multiply-conjugated metabolites are eliminated in urine rather than bile, probably because they are not substrates for the canalicular transporters. For UDCA, conjugation with N-acetylglucosamine is a very minor pathway—less than 0.3% of the oral dose.

Glutathione conjugates of bile acids (presumably as thio-esters) have been identified in rat and human infant bile by the group of Shigeo Ikegawa, but their concentration is several orders of magnitude less than that of the major biliary bile acids (179).

To date all C27 bile acids have been found to be conjugated exclusively with taurine. Bile alcohols are esterified with sulfate, and thus have a pKₐ similar to that of taurine conjugates.

E. Structure-function relationships.

In the steady state adult animal, cholesterol input – from de novo synthesis and dietary cholesterol – must be balanced by excretion – cholesterol as such and bile acids. There is thus a mole fraction of total cholesterol excretion composed of bile acids, fecal acidic steroids, divided by total (neutral plus acidic) steroids. In man, based on balance studies, this figure is 0.46; in the mouse it is 0.45, whereas in the
hamster it is 0.57, according to John Dietschy and Steve Turley (180). The very low cholesterol proportions in the biliary lipids of many vertebrate species (181), suggests that in most vertebrates this number should be well above 0.5. The route of bile acid excretion is solely via bile, as the amount of bile acids in urine is negligible. In patients with complete biliary obstruction, fecal bile acid output is extremely low, indicating that possible bile acid extruders in the small intestine epithelium are not upregulated (182). In patients with biliary obstruction, bile acid elimination is by urinary excretion, albeit with greatly increased levels in plasma and liver. In addition to cholesterol excretion in bile (as such and after conversion to bile acids) there is a small flux of cholesterol directly from the small intestine (183).

For a bile acid to serve as the chemical form of cholesterol elimination there are only a few molecular requirements. The molecule must be too polar to be absorbed passively and too large to pass through the paracellular junctions between enterocytes. It also must be resistant to pancreatic enzymes.

The bile acid molecule is too large to pass via the tight junctions of the intestinal epithelium. Amidation with glycine or taurine converts unconjugated bile acids from weak acids (pKₐ of 5) to a pKₐ of 4 for glycine amidates and a pKₐ of less than 2 for taurine conjugates (184). Thus, bile acids are present in bile and intestinal content as organic anions and therefore will not cross membranes unless a transporter is present. Bile acid glycine and taurine amidates are resistant to pancreatic carboxypeptidases, whereas bile acids amidated with amino acids other than glycine or taurine (with the exception of aspartic acid and cysteic acid), are hydrolyzed rapidly by pancreatic carboxypeptidases (185).

In vertebrates with a cecum, deconjugation and dehydroxylation convert bile acids to membrane permeable molecules and they are in part absorbed, as signaled by the presence of unconjugated bile acids in the systemic circulation. Isao Makino, who had worked in the Sjövall laboratory in Stockholm and returned to Japan, was among the first to report the presence in healthy individuals of unconjugated bile acids in plasma using gas chromatography (186). Some decades later, this early finding was confirmed by Kenneth Setchell et al, using gas chromatography-mass spectrometry (187). Nonetheless, absorption
from the cecum is not complete, as unconjugated bile acids have poor solubility at the acidic pH (pH 5.5) of the cecum and are bound to bacteria and unabsorbed dietary components.

Unconjugated (and conjugated) bile acids that are not absorbed are finally eliminated from the body by the process of defecation. As there is negligible excretion of bile acids in urine, breath, and skin, and as the bile acid nucleus is considered to be indestructible by bacterial enzymes, fecal excretion of bile acids is equal to their synthesis from cholesterol. However, it must be stressed that there is little information on the fecal bile acid composition across the vertebrate spectrum (188). There are some species that by our analyses have very low fecal bile acids, suggesting that the bile acid molecule can in fact be degraded by bacteria. The degradation of bile acids by individual bacterial strains with bile acids as the major carbon source has been characterized in great detail by S. Hayakawa (189). The biological relevance of these studies is unclear. To date, there are no studies in which bile acids tagged with $^{14}\text{C}$ in the steroid nucleus are administered and $^{14}\text{CO}_2$ in breath is collected.

There is very little information on the fate of bile alcohols in the distal intestine. In the Asiatic carp we found unchanged $5\alpha$-cyprinol sulfate in the distal intestine (190). Even if there were hydrolysis of the ester sulfate, the liberated bile alcohol is likely to have negligible passive membrane permeability because of the number of hydroxyl groups.

For solubilizing lipids in the form of mixed micelles, bile acids must be amphipathic, i.e. they must have a polar side and a non-polar side. So far as is known, all major natural biliary bile acids and bile alcohols are amphipathic, although the hydrophobic side (beta face) is somewhat reduced in UDCA. **Figure 5** shows a space-filling model of a bile acid molecule, emphasizing its amphipathic nature. To the best of our knowledge, no one has synthetized a molecule unrelated to the bile acid molecule which has the same physicochemical properties of bile acids as regards solubilizing membrane lipids.

### III. Measurement

#### A. Isolation and chromatographic separation
Elucidation of the metabolism of bile acids in health and disease requires accurate, reliable, and sensitive methods for measurement, as well as techniques for extracting bile acids (or alcohols) from complex biological matrices. The complexity of bile acid structure and the enormous range in polarity of bile acids and their conjugates have always proved a challenge to the analytical chemist. A cogent review by W. Griffiths and J. Sjövall on current analytical approaches was published in 2010 (9). These two chemists have contributed greatly to the advances in bile acid analysis.

The modern era of bile acid analysis began with the development of improved chromatographic methods for separation. Paper chromatography based on liquid-liquid partition was developed in the 1940’s by the great English chemists A.J. P. Martin and R.L. Synge who shared the Nobel Prize in 1952. In the 1950’s it was first applied to bile acids by Jan Sjövall to separate individual unconjugated bile acids (191). Sjövall and Haslewood then collaborated to apply the method to Haslewood’s vast collection of bile samples from different animals (192).

Gas-liquid (partition) chromatography was developed some 10 years later by A.J.P Martin working with A.T. James. The earliest application to bile acids came in the 1960’s (193,194), and Eneroth and Sjövall tabulated retention times for different stationery phases in 1971 (195). A decade later capillary gas chromatography was described, greatly increasing resolving power. Gas chromatography using flame ionization detectors affords both separation and quantification. However, only when a peak is further analyzed by mass spectrometry can one be one certain that the peak is truly a bile acid.

Thin-layer adsorption chromatography (TLC) became widely used when a commercial device for coating the adsorbent (usually silicic acid or alumina) on glass plates became available and excellent adsorbents were developed (196). TLC offered useful separations of bile acids by class (taurine conjugates, glycine conjugates, unconjugates) but could not separate individual taurine dihydroxy-conjugates from each other. For unconjugated bile acids, molecules separate according to the number and orientation of hydroxyl groups, but a clear separation of DCA from CDCA with the commonly employed
solvent systems does not occur. Moreover, quantification of individual spots was difficult. Nonetheless, TLC is a highly useful quasi-quantitative method for assessing purity of bile acids and monitoring chemical reactions.

The senior author was privileged to have access to the first TLC apparatus to enter Sweden, purchased in 1960 by Borgström for the separation of lipids. Borgström had been impressed with the remarkable separations of neutral lipids that were obtained using this technique by Malins and Mangold (197), then working at the Hormel Institute in Minnesota. With TLC using silicic acid as the stationery phase and phosphomolybdic acid (in ethanol) as the detection reagent most bile acids (as well as unsaturated lipids) appear as deep blue spots on a brilliant yellow background (the colors of the Swedish flag). A technique was described for coating silicic acid on microscope slides (198), and this in turn permitted screening of multiple solvent systems for optimal bile acid separation. Systems were developed (199) that our laboratory still uses today.

Reversed phase TLC has also been used for bile acids (200,201), but has not achieved popularity because of the much lower capacity of the method, as well as difficulties in detecting bile acids after separation. Eneroth and Sjövall tabulated Rf values using TLC for a large spectrum of bile acids and a number of solvent systems in their review in 1971 (195). Some of their values have been shown as actual simulated chromatograms in a chapter by the senior author (202).

In our laboratory, we have found TLC to be of great value for quantifying hepatic biotransformation of radioactive bile acids. Small (1 mm) bands of adsorbent can be directed into scintillation counter vials, using the automatic device described by Snyder and Kimble (203), or manually, by transferring the adsorbent to a vial using a brush.

HPLC was developed in the 1970’s and soon applied to bile acids. Today, the technique when used for bile acid analysis is usually executed in the reversed phase mode, i.e. with C_{18} or C_{8} bonded silicic acid as the lipophilic solid phase. Numerous laboratories have reported the power of this analytical method for
separating and measuring conjugated bile acids in bile and intestinal content (9). Conjugated, i.e. N-acyl amidated bile acids have an amide bond that has absorbance at 205 nm, permitting their measurement in the effluent. Unconjugated bile acids, non-amidated bile acid sulfates, and bile alcohol sulfates are not detected at 205 nm. HPLC with UV detection is too insensitive for analysis of serum bile acids unless large volumes of serum are used (204). One solution to the non-detection of unconjugated bile acids and bile alcohol sulfates is to use a light scattering detector that responds approximately equally to all bile acids (205). A second approach, useful for unconjugated bile acids, is to couple the bile acid to a chromophore such as the phenylacyl group (206).

In our laboratory, we developed an HPLC method for separating the 12 conjugated bile acids predominant in human bile (165), and we have found this method to be essential in our analysis of vertebrate bile samples (82). Nonetheless, when used alone for bile analysis, there are many unknown peaks that may or may not be bile acids. Therefore peak identification requires confirmation by mass spectrometry.

B. Mass spectrometry

Mass spectrometry had been developed in the 1930’s and was used to analyze mixtures of small molecules. Mass spectrometry of large organic molecules emerging from the gas chromatograph was developed by R. Ryhage in Sweden, and the first commercial instrument for this purpose was marketed by LKB in 1965. William Elliott wrote in 1988, “Two years later [1965], I returned to the Karolinska to devote the summer to learning about the GC/MS which Ragner Ryhage had built, and was then produced by LKB. Our LKB 9000 is still operating, beginning its 23rd year.” Since then, there have been continuous improvements in techniques for sample volatilization and instrument sensitivity, culminating in the MALDI analysis of proteins. In our opinion, anyone who wishes to do a complete bile acid analysis must have HPLC/MS/MS available. It is also desirable to have GC/MS availability when needed, as capillary gas chromatography has greater resolving power than HPLC.
The wedding of HPLC and MS/MS is thus an essential step in the analysis of unconjugated and conjugated bile acids. Currently, internal standards containing three or more deuterium atoms are not available for many bile acids and are sorely needed. For a total analysis, class separation by ion exchange chromatography developed by Setchell and Sjövall should be performed before HPLC/MS/MS (9). For state of the art analysis, either capillary GC or capillary HPLC will be required to separate complex mixtures of bile acids. For unknown bile acids, structure assignment requires NMR for elucidation of structure, and ultimately synthesis for confirmation of assigned structure.

C. Enzymatic and competitive binding assays

Development of an enzymatic technique for determining total $3\alpha$-hydroxy bile acids based on oxidation of the $3\alpha$-hydroxy group by a $3\alpha$-hydroxy steroid dehydrogenase was a major advance in bile acid analysis (207), especially for workers in the field of bile acid physiology. This technique was based on the pioneering work of Talalay (208), who developed enzymatic analysis of $3\alpha$-hydroxy hormonal steroids using a steroid dehydrogenase isolated from Pseudomonas testosteroni. Details of the method were clarified by Turley and Dietschy (209). The sensitivity of the method was increased by measuring the reduced NAD by changes in fluorescence or by coupling the oxidation to reduction of a tetrazolium dye (210). The enzyme technique using $3\alpha$-hydroxysteroid dehydrogenase does not measure $3\beta$-hydroxy bile acids. Thus, when used for determination of fecal bile acids, it will underestimate total bile acids (119). An enzyme recycling system has recently been marketed by Diazyme™ and is quite sensitive.

In our laboratory, based on collaboration with Marlene DeLuca. Aldo Roda developed an enzymatic bioluminescence assay for primary bile acids. The method was based on co-immobilizing on beads a $7\alpha$-hydroxysteroid dehydrogenase, an oxido-reductase, and luciferase (211). The method was validated by showing excellent agreement with GC/MS measurements (212). However, this method, despite its simplicity and sensitivity, has not been widely adopted. Limitations in the enzymatic method are summarized in the review by Griffiths and Sjövall (9).
All of the chromatographic techniques discussed above are labor-intensive and time-consuming. For measuring diurnal variations in plasma levels of total bile acids, two approaches have been taken to develop a rapid, simple procedure for measuring total bile acids. The first was the enzyme method that measures total $3\alpha$-hydroxy bile acids, as just discussed. The second approach was the development of competitive binding techniques, such as radioimmunoassay. What is measured depends on the affinity of the antibody, and thus one determines “immunoreactive” bile acids. At the Mayo Clinic, Wilfred Simmonds et al developed a radioimmunoassay specific for choly conjugates (213) which was then used by LaRusso et al to define the time course of the postprandial elevation of such choly conjugates in health and in patients with bile acid malabsorption (214). Later, Schalm et al (215) developed a radioimmunoassay for chenodeoxycholy conjugates, and used it to show that the postprandial peak of CDC conjugates occurs earlier than that of choly conjugates (216), as shown in Figure 6. Other radioimmunoassays have been developed, as summarized in the review by Griffiths and Sjövall (9).

All of the above methods for measuring bile acids isolate the bile acids from the body tissue or fluid, and then quantify the bile acids. We have long needed a bile acid-sensitive molecule whose signals would provide information on the concentration of bile acids within the living cell. This need has now been supplied by the development of an indicator molecule consisting of a protein possessing a bile acid binding site (based on the binding site of FXR) and two reporter molecules whose fluorescence is determined by the extent to which bile acids occupy the binding site -- the greater the binding, the less the fluorescence. The technique, as described by van der Velden et al (217), uses the well-known Forster Resonance Energy Transfer (FRET) principle, and has already been used to characterize bile acid transport into isolated hepatocytes.

D. Diagnostic utility of plasma bile acid levels
Despite the simplicity of competitive binding techniques for measuring conjugated bile acids, they have not been widely adapted by the clinical chemists. The measurement of plasma bile acids is not part of the usual panel of liver tests (aminotransferases, alkaline phosphatase, total bilirubin). The major reason is that bile acid levels are probably inferior in sensitivity and specificity in detecting liver injury when compared to the standard liver panel. This conclusion is based on clinical studies (218,219). In addition, a recent animal study (220) in which liver injury was experimentally induced in rats by feeding a variety of toxic molecules found that serum bile acid levels were not more sensitive than aminotransferase levels. Thus, because of the lack of greater sensitivity of serum bile acid measurements, there is the belief that the current liver tests are adequate. Moreover, the instruments that automate determination of a serum “chemical panel” do not have a bile acid “channel”. For purely cholestatic diseases such as primary cholestasis of pregnancy, the level of serum bile acids has value for predicting prognosis (221).

**E. Determinants of bile acid profiles in body fluids**

In assessing the significance of bile acid concentrations that in turn define the bile acid profile, it is important to realize that plasma, hepatic, biliary, urinary and fecal bile acids all have a different composition in health. Plasma bile acids, in health, represent the balance between instantaneous intestinal input and net hepatic uptake, with first pass clearance ranging from 40 to 90%, depending on the bile acid. Hepatic bile acids (as distinguished from hepatocyte bile acids) represent the balance between input of unconjugated and conjugated bile acids, subsequent biotransformation of these molecules, as well as bile acids that have been secreted by transporters into the canaliculus. Urinary bile acids represent filtered bile acids plus any bile acids secreted by the tubules minus those absorbed by the tubules. Ductal bile which is usually analyzed differs from canalicular bile by that fraction of conjugated and unconjugated bile acids that are absorbed by the biliary ductules. Gallbladder bile is ductal bile minus bile acids absorbed from the gallbladder. Fecal bile acids represent bile acids entering the colon with or without modification by bacterial enzymes minus those that were absorbed.
The power of current analytical techniques may result in an enormous variety of bile acids being detected. When such results are tabulated, we suggest organizing bile acids by primary bile acids, and then listing secondary bile acids derived from the primary bile acid underneath their precursor. Bile acids that are present in trace proportions should be lumped together to simplify the presentation. The choice of the analytical method should be determined by the biological question that is being asked. For nomenclature in reporting bile acid analyses, the recommendations of an *ad hoc* nomenclature committee (117) should be followed.

**IV. Bile Acid Metabolism: transport and the enterohepatic circulation**

**A. Bile acid transport**

1. Overview

Glycine and taurine conjugated bile acids are fully ionized at the pH conditions prevailing in plasma, bile, and small intestinal content. Therefore, these molecules will not cross a cell membrane and enter a cell unless a transporter is present. What is responsible for the enterohepatic circulation is the vectorial transport of conjugated bile acids through the ileal enterocyte and the hepatocyte. The enterohepatic circulation of bile acids results from the action of two chemical pumps and three fluid flows. The chemical pumps are the vectorial transport systems of the ileal enterocyte and the hepatocyte – each of which has basolateral and apical transporters. The fluid flows are the aboral movement of intestinal content due to intestinal propulsive activity, portal venous blood flow that transports bile acids from the intestine to the liver, and biliary flow that returns the bile acids to the intestine. The molecules that constitute the circulating pool of bile acids must be substrates for the vectorial transport systems of both the ileal enterocyte and the hepatocyte in order to circulate enterohepatically.

The initial input into the hepatic limb of the enterohepatic circulation consists of primary bile acids formed and conjugated in the pericentral hepatocytes. The newly synthesized conjugated bile acids are pumped across the canalicular membrane and flow in canalicular bile from the pericentral region to the
periportal region, their direction of flow being opposite from that of sinusoidal blood. In the periportal region, newly synthesized bile acids are joined by bile acids returning from the intestine and pumped into bile.

The return pathway of the enterohepatic circulation features three inputs into the portal venous blood from bile acids that have been absorbed by enterocytes. The dominant input into portal venous blood consists of “undamaged” conjugated bile acids that have been absorbed by the ileal transport system. A second input is that of unconjugated primary bile acids formed by bacterial deconjugation in the distal small intestine and large intestine. A third input is that of “secondary (damaged) bile acids that are formed by bacterial modification of the hydroxyl groups of primary bile acids. Only a fraction of the secondary bile acids that are formed are absorbed, and those that are not absorbed are eventually eliminated by defecation.

As discussed previously, those secondary bile acids that are absorbed from the distal intestine may undergo uptake by the hepatocyte or in some cases, urinary excretion (or both). After hepatocyte uptake, “damaged” bile acids undergo “repair” in the hepatocyte by “reconjugation” as well as re-epimerization of hydroxy groups (from \( \beta \) to \( \alpha \)) and reduction of oxo groups to hydroxy groups. They are then secreted into bile to join the recycling primary bile acids. In man, the flux of bile acids returning from the intestine is at least twentyfold greater than the input of newly synthesized bile acids. More details of bacterial damage and hepatocyte repair are given in the introduction.

In man, the major secondary bile acids in colonic content are lithocholic acid, formed by 7-dehydroxylation of CDCA, and DCA, formed similarly from cholic acid. In man, lithocholic acid is not only N-acylamidated, but also sulfated in part. Such sulfo-lithocholyl amidates are secreted into bile, but are poorly absorbed from the small intestine (222), so they do not have an enterohepatic circulation. Those unsulfated amidates secreted in bile return to the liver, undergo sulfation in part, are re-secreted into bile, and eliminated via the fecal route. The end result is that lithocholic acid has a very rapid
turnover in man, and sulfated and unsulfated lithocholyl amidates constitute less than 5% of biliary bile acids (165). In rodents, lithocholic acid is not sulfated but undergoes hydroxylation at C-6 or C-7 (15).

As noted previously, DCA may pass through the hepatocyte without further metabolism, or may be rehydroxylated to varying degrees in a species dependent manner. The “repaired” and reconjugated bile acids are secreted from the hepatocyte and join the cycling conjugated bile acids.

Reabsorption of bile acids from the small intestine gives a cycling “pool” for each bile acid. The size of the bile acid pool and its turnover rate differ for each bile acid, although the turnover rate of CDCA and DCA are about the same. Most of each individual bile acid pool is stored in the gallbladder during overnight fasting, and biliary bile acid composition is determined by the relative proportions of each type of bile acids, this in turn being determined by the sizes of the individual bile acid pools.

2. Transport by the ileal enterocyte: the apical transporter

The first person to observe preferential absorption of conjugated bile acids by the ileum was Hermann Tappeiner, who presented his findings to the Vienna Academy of Sciences in 1887 (223). Tappeiner studied the absorption of taurocholate from different regions of the small intestine using anesthetized dogs. He found that taurocholate was absorbed preferentially in the ileum. Rather than concluding that the ileum had special properties, Tappeiner was puzzled by the failure of the proximal intestine to absorb taurocholate. Verzar, working in Hungary, authored one of the first monographs on intestinal absorption (224). He was skeptical of the results of Tappeiner and directed his colleague Fröhlicher to repeat Tappeiner’s experiments. Fröhlicher obtained identical results (225), again showing that ileal absorption of conjugated bile acids was greater than that of jejunal absorption. However, in Verzar’s book on intestinal absorption the concept of specific transporters with differing substrate affinities is not considered. At that time, the mechanism of absorption of small molecules was considered to be passive, analogous to diffusion.
Baker and Searle, working at the University of Iowa, rediscovered the work of Tappeiner and Fröhlicher, and confirmed the selective ileal absorption of taurocholate in the rat ileum in 1960 (226).

A decade before, T.H. Wilson, an American working in the laboratory of G. Wiseman at the University of Sheffield, developed the everted gut sack technique and used this method to show that glucose was actively transported (227). (It is rumored that this technique was the result of a drunken party in Paris some years before, where physiologists were bemoaning the difficulty of studying intestinal absorption in the intact animal. Someone, never identified, said the only solution was to turn the animal inside out. And someone else is supposed to have said, “Maybe that’s a good idea!” In 1960, Robert Crane, who had studied glucose absorption since his postdoctoral studies with the Cori’s in St. Louis, published his idea that glucose absorption was sodium dependent, and that the downhill transport of sodium served to energize the uphill transport of glucose (228). (The late Robert Crane was a tall, handsome charmer who always had an eye for attractive women. At a party, he tried to persuade a lovely young scientist to go home with him; when she refused, he started attacking her data).

In 1961, Leon Lack and Ike Weiner, young faculty members at Johns Hopkins, were discussing how to teach medical students about the enterohepatic circulation. They realized that nothing was known about the mechanism of bile acid absorption. They used the everted sac technique to show that the last fourth of the small intestine of rats and guinea pigs actively transport conjugated bile acids against a concentration gradient (229). Their data are illustrated in Figure 7. As noted earlier, the senior author, working with his mentor, Bengt Borgström and Göran Lundh (a surgeon and brother-in-law of Borgström) perfused the small intestine of healthy volunteers and obtained strong evidence for ileal absorption of conjugated bile acids in man (20). Peter Holt, working in the laboratory of Kurt Isselbacher at Harvard, showed that conjugated bile acid absorption in the everted sac preparation was sodium-dependent (230). Thus, it seemed that a Na\(^+\) dependent transporter for conjugated bile acids was present in ileal enterocytes. (The term “enterocyte” was coined by Christopher Booth, a prominent English gastroenterologist and bibliophile, who discovered that vitamin B12, like conjugated bile acids, was also absorbed.
predominantly in the ileum). In 1978, H. Löcke, G. Stange, R. Kinne and H. Murer, working at the Max-Planck Institute in Frankfurt, prepared vesicles from ileal enterocyte brush borders, and used this preparation to show that taurocholate uptake was carrier mediated and Na⁺ dependent (231). Attempts were made to isolate the apical bile conjugated bile acid transporter by G. Kurz, working in Freiburg, using the photoaffinity technique with ³H-tagged azido-bile acids (232). These studies were continued by his student, Werner Kramer at the Hoechst Company in Frankfurt (233), but they failed to clearly identify the ileal apical bile acid transporter.

During this time, ideas of cell polarity were developing (234). Thus, it seemed that at a minimum, four transporters would be required for the enterohepatic circulation of bile acids. These would be an apical and basolateral transporter in the ileal enterocyte, and an apical (canalicular) and basolateral (sinusoidal) transporter in the hepatocyte.

The great breakthrough, in our judgment, came in 1987 with the use of expression cloning by Matthias Hediger working in the laboratory of Ernie Wright at UCLA. Using the technique of injecting mRNA into frog eggs (obtained from *Xenopus laevis*), they were able to report the cloning of the sodium coupled transporter for glucose (235). Seven years later, Paul Dawson decided to attempt to clone the ileal bile acid transporter. Dawson was well educated for the task, having worked with Joseph Goldstein, Michael Brown, and David Russell at UT Southwestern in Dallas. Dawson, now relocated to Wake Forest University, reported the successful cloning of the ileal apical sodium-dependent bile acid transporter (protein, ASBT; gene, *SLC10A2*), using a somewhat different expression technique (236). This achievement, more than a century after Tappeiner’s original study, permitted the testing of individuals for defects in the gene encoding ASBT, as well as providing a molecular target for pharmaceutical companies who might wish to develop an inhibitor.

In Dawson’s original paper, he found that ASBT was also present in the kidney, in agreement with earlier studies by F. Wilson et al (237). Later, the laboratories of G. Alpini in Texas (238) and N. LaRusso
(239) at the Mayo Clinic showed that ASBT was also present in cholangiocytes. ASBT is highly specific for bile acids, although the integrity of the 7α-hydroxy group seems more important than that of the 3α-hydroxyl group, as some substituents can be added to the 3α-position without impairing transport (240). As yet, the transporter has not been crystallized, although the structure of a bacterial homologue has recently been published (241). That paper proposes that two sodium ions act allosterically to open a binding pocket. An overview of the SLC10 transporter family is available (242).

3. Transport by the hepatocyte: bile acid uptake

Bile acid uptake by the liver was known to be highly efficient. For example, in an isolated perfused dog preparation studied at Mayo Clinic by N. Hoffman et al, uptake of taurocholate was 100% and that of the taurine conjugate of CDCA was 70% (243). In a study published that same year from the Erlinger laboratory in Paris, first pass extraction of taurocholate in the dog was found to be 40-80% and saturable, consistent with uptake being carrier mediated (244). Conjugated bile acid uptake was saturable, and using either hepatocytes (245) or the isolated perfused rat liver (245) was sodium-dependent (246). Thus it seemed highly likely that one or more sodium-coupled transporters was present in the sinusoidal membrane and mediated the efficient uptake of bile acids.

A sodium-dependent basolateral hepatocyte protein, named NTCP (sodium taurocholate porter) was cloned in 1991, again using expression cloning, by Bruno Hagenbuch, Bruno Stieger and colleagues working in the laboratory of Peter Meier in Zurich (247). When the Na⁺ coupled ileal transporter ASBT was cloned four years later, it was possible to compare the sequences of NTCP and ASBT and identify some regions of homology (248). Thus, the ileal uptake protein and the hepatocyte removal protein were related, and both used the sodium gradient to power the uptake. However, NTCP has a much broader specificity than ASBT (249,250). NTCP can be considered an anion transporter that transports not only bile acids but also other organic anions into the hepatocyte from plasma, whereas ASBT selectively conserves solely conjugated bile acids, thereby generating their enterohepatic circulation.
NTCP is not the only bile acid transporter in the basolateral (sinusoidal) membranes of the hepatocyte. A family of Na+ independent transporters, organic anion transport polypeptide (OATP) was identified, and several of these multi-specific transporters have the ability to transport bile acids, as tabulated in cogent reviews by Hagenbuch and Stieger (249) and Klaassen and Alaskans (250). In man, these include OATP1A2, OATP1B1, OATP1B3, and OATP1C1. The genes encoding OATP transporters are denoted by the term solute carrier organic anion (SLCO).

To what extent OATP transporters contribute to the highly efficient uptake of conjugated bile acids from sinusoidal plasma is not known, but it is believed that the majority of bile acid uptake is mediated by hepatocyte NTCP. The efficient uptake of dihydroxy conjugated bile acids is truly remarkable, as these are 99% bound to albumin (251), and presumably uptake is from the extremely low concentration (nM) of unbound species. In our view, it is reasonable to think that more than one transporter may be involved in the sinusoidal uptake of conjugated bile acids. Presumably the transporters involved in bile acid uptake have a zonal distribution and are enriched in periportal hepatocytes.

During enterohepatic cycling, there is continuous bacterial deconjugation in the small intestine followed by hepatic reconjugation, presumably occurring in the periportal hepatocytes. This was first recognized in rat studies by Norman, working in Lund, Sweden in 1955 (252) and much later confirmed for man in our own studies at the Mayo (253). The rate of “reconjugation” in the periportal cells is several fold greater than the rate of conjugation of newly synthesized bile acids in the pericentral cells.

The uptake of unconjugated bile acids is by both passive and carrier-mediated mechanisms, and depends on the number and orientation of hydroxyl substituents. Lithocholic acid, CDCA, DCA, UDCA, and presumably, 3,6-dihydroxy bile acids are highly membrane permeable (221,245) and in our judgment do not require a transporter for rapid uptake into the hepatocyte. Cholic acid, in contrast, passes across cell membranes slowly and requires a transporter for rapid cellular entry. A systematic study of the passive membrane permeability of the three muricholic acid epimers is lacking.
At least two transporters appear to be involved in the uptake of unconjugated trihydroxy-bile acids. The laboratory of Curt Klaassen in Kansas City showed that OATP1B2 (a sodium-independent transporter), appears to be involved in the uptake of unconjugated bile acids that return from the intestine (254). Knockout of the transporter in mice caused a marked elevation in the plasma level of unconjugated bile acids. Hubbard et al (255) showed that fatty acid transport protein 5 (FATP5) knockout mice have a major fraction of unconjugated bile acids in bile, and such unconjugated bile acids were enriched in secondary bile acids. They suggested that FATP5 was involved in both the uptake and CoA formation of unconjugated bile acids returning from the intestine. It is attractive to postulate that this transporter mediates not only adenylate formation but also subsequent CoA thioester formation of the unconjugated bile acids that enter the hepatocyte. Steinberg et al (256) characterized a homolog of a peroxisomal very long chain CoA synthetase. This homolog was present in the smooth endoplasmic reticulum. They found that it had CoA synthetase activity for cholic acid, and postulated a role for this enzyme in cholic acid reconjugation.

In the isolated perfused liver, rapid uptake of DCA occurs, presumably followed immediately by CoA formation. Formation of the adenylate and subsequent CoA thioester serves to prevent back diffusion across the sinusoidal membrane and thereby traps the bile acid CoA in the hepatocyte. In contrast, 24-norDCA, which does not form a CoA thioester, regurgitates back from the hepatocyte into the perfusate (257).

The basolateral transporter multidrug resistant protein 4 (MRP4) has been shown to mediate the co-efflux of conjugated bile acids and reduced glutathione, in work from the laboratory of Dieter Keppler (258). For this work, we provided tritium-labeled conjugated bile acids of high specific activity (259). Keppler has suggested that conjugated bile acids percolate in and out of the sinusoidal membrane, and that such efflux from the hepatocyte into the space of Disse may serve to prevent toxic concentrations of bile acids in the hepatocyte (260).
4. Canalicular secretion of bile acids

The next challenge was to identify the canalicular transporter or transporters. This had to be a powerful transporter, because the intracellular concentration of bile acids was thought to be in the low µM range, whereas the concentration of monomeric bile acids in ductal bile, based on the dialysis studies of Duane (261) was about 1000 µM. There was thus an uphill transport of three orders of magnitude.

To study this transport system, one needed vesicles consisting predominantly of canalicular membranes. The first preparation was achieved by Inoue and Kinne, then working in the laboratory of Irwin Arias in New York (262). The canalicular membranes were isolated by nitrogen cavitation and calcium precipitation. At about the same time, Peter Meier working in the laboratory of James Boyer at Yale developed a method for preparing canalicular vesicles using zonal flotation and high speed centrifugation through discontinuous sucrose gradients (263). These vesicles were used by Meier and his colleagues to show that taurocholate uptake by canalicular vesicles was carrier mediated and probably electrogenic (264). A similar study was published by Inoue, Kinne, and the Arias group (265).

In 1986, Eamon O’Maille, who had studied bile acid secretion in the intact dog, was working in our laboratory as a guest investigator. He carefully measured the $T_{\text{max}}$ of taurocholate transport in the rat and concluded it was too great to be explained by membrane potential alone (266). Y. Adachi returned to Japan from the Arias lab and in 1991 reported that taurocholate was actively transported into canalicular vesicles if an ATP generating system was present (267). That same year, the Arias lab (Nishida et al) also reported that rat canalicular membrane vesicles contain an ATP-dependent bile acid transport system (268).

A year or so later, the Zurich group began to work on the ATP-dependent process of conjugated bile acid extrusion by the canaliculus. They studied a variety of cell systems to find one which did not display ATP-stimulated bile acid transport and finally ended up with Sf9 (insect) cells. They noted a paper from the laboratory of Victor Ling in Vancouver, Canada that showed a partial sequence of a transport protein
named “sister of P-glycoprotein,” later named multidrug resistance 1 (MDR1) (269). Using this sequence, an antibody to the putative canalicular transporter was generated. The Zurich group then used phage libraries, their antibody, and Sf9 cells to monitor expression. They were ultimately successful in cloning the bile salt export pump (BSEP) (protein, bile salt export pump; gene, \textit{Abc11A1}) (270). Our laboratory provided \textsuperscript{3}H-labeled bile acids with high specific activity (259) that were used to define its substrate affinity.

It was not long until children with cholestatic liver disease were found to have mutations in \textit{Abcb11A1}, the gene encoding BSEP (271). This disease had already been named Progressive Familial Intrahepatic cholestasis, type 2 (PFIC2), and as the defect is canalicular, markers of ductular injury such as $\gamma$-glutamyl-transferase are not elevated in patients with this genetic defect. The name PFIC2 is widely used, but it seems preferable to call the disease BSEP deficiency. There are other ABC-type canalicular transporters such as MRP2, MDR1, and BCRP but these are not believed to be very important in the canalicular transport of unsulfated bile acids in man, based on the severe cholestasis that occurs in patients with BSEP deficiency as well as evidence from in vitro studies of MRP2 (272).

The laboratory of Victor Ling, in Vancouver, reported the creation of an \textit{Abcb11} knockout mouse, which appeared to have only mild cholestasis. (273). The mildness of the mouse phenotype contrasted strikingly with the severe cholestatic phenotype that occurs in infants with a genetic defect. This striking difference between BSEP deficiency in humans and that in mice appears to have at least three explanations. First, in mice, a major bile acid is $\beta$-muricholic acid, which is much less cytotoxic than DCA and CDCA, these being major bile acids in man. Second, in mice, MDR1 is upregulated and is likely to be an effective conjugated bile acid extruder (274). Third, BSEP deficient mice form tetra-hydroxy and penta-hydroxy bile acids (275,276), some or all of which are likely to be transported by MRP2 (277). Humans with cholestatic liver disease do not form tetra- or penta-hydroxy bile acids to any extent. In mice, the defect in bile acid export from the hepatocyte is easily shown by feeding such mice
cholic acid (278), which in this condition is highly toxic. Impaired β-fatty acid oxidation is also present in such cholic acid-fed animals (279).

5. The basolateral transporter of the ileal enterocyte

There was one last carrier to be identified, the basolateral transporter of the ileal enterocyte. Also, if conjugated bile acids were absorbed by cholangiocytes using ASBT and returned via the periductular capillary plexus to the hepatocyte, there had to be a basolateral transporter in cholangiocytes. The hepatocyte did not require a basolateral efflux transporter, as it already had one in MRP4, and most bile acid efflux from the hepatocyte was across the canalicular membrane into bile mediated by BSEP.

In 2001, Wei Wang working with James Boyer and Nazzareno Ballatori at the Mount Desert Island identified two genes encoding a Na+ independent bile alcohol sulfate transporter in the skate (280). The genes were isolated using the *Xenopus laevis* technique of expression cloning, and they mediated transport only when they were expressed simultaneously. They were named organic solute transporters A and B (OSTα/OSTβ). Subsequently, Ballatori collaborated with the laboratory of Paul Dawson and identified this dimeric organic solute transporter as the long sought ileal enterocyte basolateral transporter (281). The two proteins were present predominantly in ileum and kidney, and immunohistochemistry showed that the dimeric transporter was present on the basolateral membrane of the ileal enterocyte and renal proximal tubular cell. When OSTα/OSTβ was expressed in polarized monolayers at the basolateral domain and ASBT was expressed at the apical domain, bile acid transport from the apical side to the basolateral side was observed (281). Further study of OSTα/OSTβ showed that it was a facilitated diffusion carrier and not entirely specific for bile acids, in contrast to ASBT, which is highly specific for bile acids. In cholestasis, hepatocyte OSTα/OSTβ is upregulated on the basolateral (sinusoidal) membranes, and mediates bile acid efflux from the hepatocyte into sinusoidal plasma (282).

To date, no diseases caused by defective OSTα/OSTβ have been identified. A mouse knockout has been created, and it differs from the ASBT knockout in that with the OSTα/OSTβ knockout, there is no
marked increase in bile acid biosynthesis despite bile acid malabsorption (283). In contrast, in the Slc10A2 (ASBT) knockout, bile acid synthesis increases nearly twentyfold (284). This apparently anomalous absence of a compensatory increase in bile acid synthesis is presumed to be the result of bile acids accumulating in the ileal enterocyte and upregulating the nuclear receptor FXR, which in turn stimulates the release of FGF15 into portal venous blood (285). Hepatocyte uptake of FGF15 acts to downregulate bile acid synthesis, as noted in the Introduction. If there is no absorption of bile acids from the ileum, yet synthesis does not increase, the net result should be a bile acid deficiency in the proximal small intestine, as secretion will fall and become equal to synthesis. This in turn should cause malabsorption of saturated fatty acids and fat soluble vitamins.

6. Imaging of bile acid transport

An 11C-labeled analogue of taurocholate (cholyltaurine) has been synthesized by the PET (positron emission tomography) scan center of Aarhus University Hospital (286). The molecule, choly-N-11CH3 – glycine (cholylsarcosine) permits the kinetics of conjugated bile acid hepatic uptake and excretion into the biliary tract to be measured and visualized in real time. Because of the extremely short half-life of 11C (20 minutes), this compound can be used only in PET centers that have a reactor (needed for the production of 11C) close to the patient. The clinical utility of this molecule in assessing cholestatic disease is under active investigation.

While this review was being written, a book entitled, “Hepatobiliary Transport in Health and Disease” was published (287). It should cover the topics discussed in our review in much more detail.

B. The enterohepatic circulation in health

1. History

The enterohepatic cycling of bile was clearly described by Borelli in 1681 (288). He wrote, “Bile is collected in the portal vein [from the intestine] and separated from the blood again in the liver. It is then
sent into the bile vessels. It leaves these and repeats its circuit thought the intestine and the mesenteric veins again and again.” Liebig, a prominent German chemist who worked in the mid-19th century, was fully aware of the enterohepatic circulation of some or all of the constituents of bile (289). Lehmann in his text book of physiological chemistry, published in 1853, wrote,”the greater part of the bile is again resorbed as it passes through the intestine. The chief biliary constituents which are resorbed are the soluble salts and cholic acid” (290). Moritz Schiff, working in Florence, Italy, prepared anesthetized dogs with an “amphibolic” catheter. When bile was diverted to the outside, bile flow stopped; when bile was allowed to enter the intestine, bile flow continued. Schiff concluded that bile secretion requires the intestinal absorption of bile acids (291). (Schiff was put on trial by the English in Florence for unnecessary cruelty to animals, as he used anesthetized dogs in his experiments. He was found guilty, and forced to leave Italy. Schiff established his laboratory in Geneva, Switzerland, and became a leading physiologist in Europe in the 19th century). Additional information on the history of knowledge about the enterohepatic circulation is given in the detailed review of the enterohepatic circulation by M. Carey and W. Duane published in 1994 (292).

The enterohepatic circulation of bile acids denotes a mass of bile acids that is discharged by gallbladder contraction into the small intestine during digestion, is absorbed from the distal intestine, returns to the liver, and may either be stored in the gallbladder or resecreted into the intestine. When a meal is ingested, the hormone cholecystokinin is released from the small intestine. Cholecystokinin induces gallbladder contraction as well as relaxation of the valve (sphincter of Oddi) at the end of the common bile duct where it empties into the small intestine. Bile then enters the duodenum. Some of the bile acids are absorbed in the jejunum, but most are transported by intestinal peristalsis to the distal ileum where they are efficiently absorbed. The bile acid molecules pass through ileal enterocytes, and enter portal venous blood to return to the liver. One of the early illustrations of the enterohepatic circulation of bile acids with values for man was presented by Sune Bergström in 1959 (8) and is shown in Figure 8. Whether ancient
vertebrates, e.g. the coelacanth and the hagfish, have an enterohepatic circulation with their C_{27} bile alcohol sulfates is not known.

2. Determinants, properties and measurement of the enterohepatic circulation

As noted previously, the enterohepatic circulation is the resultant of vectorial transport by the ileal enterocyte and the hepatocyte, as well the flow in the portal venous system, the small intestine, and the biliary tract. The only molecules that have an enterohepatic circulation are those that are transported by the membrane transporters of the ileal enterocyte and the hepatocyte. Because only bile acids are transported by specific transporters, only bile acids have an enterohepatic circulation. The possible enterohepatic circulation of drugs is beyond the scope of this review.

The vast majority of the circulating bile acids are in the biliary tract and intestine, because plasma and hepatocyte levels are very low. Plasma levels are low (< 10µM) because of efficient hepatic extraction as well as dilution. Urinary levels are also low (< 10µM) because only a fraction of plasma bile acids enter the glomerular filtrate, as bile acids are bound to plasma proteins, mostly to albumin (293). Moreover, much of the fraction of bile acids that enters the glomerular filtrate is reabsorbed in the proximal tubules via ASBT and OSTα/β. We found that cholyl conjugates were 70% bound, and dihydroxy conjugates were >95% bound in human serum (251), so the glomerular filtrate should contain mostly conjugates of cholic acid in man. The end result is that the concentration of bile acids in urine in healthy man is extremely low, and urinary bile acids constitute a negligible fraction of bile acid excretion.

The mass of cycling bile acids in man was first measured by Sven Lindstedt, a member of the Bergström group, then in Lund (14). Lindstedt prepared cholic acid tagged with tritium and followed the decay of bile acid specific activity (in bile) with time in human subjects. The specific activity of the isolated cholic acid fell exponentially with time, indicating that loss of radioactivity followed first order kinetics. The slope of the natural logarithm of the bile acid specific activity decay curve can be extrapolated to zero time to permit calculation of the pool size. The slope of the specific activity decay
curve times the pool size gives the synthesis rate. The technique of isotope dilution described by Lindstedt has been used to measure the pool size and synthesis rate of the two primary bile acids – cholic and CDCA repeatedly (294). Such studies have shown that the rate of cholic acid synthesis is about twice that of CDCA.

The technique can also be performed with stable isotopes, measuring the atoms % excess in plasma bile acids by mass spectrometry, approach developed by Franz Stellaard and his colleagues in Groningen (295). When radioactive bile acids are used in the usual amounts, the specific activity of plasma bile acids is too low to measure accurately.

The isotope dilution technique has also been used for DCA (296), but input now represents absorption of newly formed DCA rather than synthesis. The input of DCA is some fraction of cholic acid synthesis, and could therefore vary from 0 to nearly 1.0. This fraction – DCA input/cholic acid synthesis has been termed the $f_{\text{dehydrox}}$ and ranges in healthy adults from 0.2 to 0.5 (297).

In the past, there has been concern because values for bile acid synthesis obtained by isotope dilution were often somewhat higher than those obtained by analysis of fecal bile acids. The belief at present, is that the discrepancy between these two methods, which should give identical results, is attributable to the difficulty of analyzing all fecal bile acids. Thus some fecal bile acids are not measured, and a higher result for bile acid synthesis is obtained by the isotope dilution technique. Also, if a considerable volume of bile is removed during the daily duodenal bile sampling used in the isotope dilution technique, this will increase bile acid synthesis.

No isotope dilution studies have been performed with C27 bile acids or bile alcohols.

3. Synthesis of labeled bile acids for metabolic studies

In the laboratory of the senior author at the Mayo Clinic, better radioactive tags for bile acids were sought in order to characterize bile acid metabolism. The Bergström group was the first to synthesize 24-
\(^{14}\text{C}\) tagged bile acids (298), and this label was stable during enterohepatic cycling. In order to prepare \(^{3}\text{H}\) tagged bile acids that would also be stable during enterohepatic cycling, we used a method developed by Peter Klein to exchange \(^{3}\text{H}\) into the C-2 and C-4 positions (299). N. LaRusso, as his first research project, examined the stability of this label, and found it was good, but not perfect (300). For labeling CDCA and lithocholic acid, we synthesized their \(\Delta^{11}\) derivatives, and reduced them with \(^{3}\text{H}\). Again, this label was good, but not perfect (301). Finally, in San Diego, in work led by our chemist, Claudio Schteingart, we prepared bile acids with a double bond between C-22 and C-23, and then reductively tritiated these molecules (259). The 22-23 \(^{3}\text{H}\) label was shown to be completely stable during enterohepatic cycling in a collaborative study performed with W. Duane at the University of Minnesota (302). Isotope dilution studies have not been performed with \(\beta\)-muricholic acid because of the non-availability of labeled material. However, the \(\Delta^{22}\) compound, the precursor for reductive tritiation, has been synthesized by the Iida group (137).

4. The enterohepatic circulation in man

Because of bacterial deconjugation in the intestine followed by absorption from the intestine and reconjugation in the liver, the turnover rate of the steroid moiety of \(C_{24}\) bile acids is considerably slower than that of its glycine moiety. As noted previously, this deconjugation-reconjugation pathway was identified in Sweden by A. Norman in rats (252), and also somewhat later by clinical studies in the laboratory of the senior author at the Mayo Clinic (253). Deconjugation-reconjugation also occurs with taurine conjugated bile acids, but the difference in turnover rates between the steroid and amino acid moiety is less, because taurine conjugates are deconjugated more slowly than glycine conjugates (162). These studies, performed at the Mayo Clinic in the laboratory of the senior author, have never been repeated, so confirmation of their findings is desirable.

No information is available on the turnover rates of either the steroid moiety or the taurine moiety of \(C_{27}\) bile acids. For most \(C_{27}\) bile alcohols, the process of deconjugation – reconjugation is unlikely, as the
liberated bile alcohol, in contrast to that of unconjugated bile acids, is too polar to be passively absorbed (190).

Early workers measured the bile acid pool by isotope dilution and guessed at the number of times that the pool cycled, the product being bile acid secretion. In our view, this assumption is not valid, and is analogous to trying to predict cardiac output from the measurement of blood volume. A flow can only be estimated by indicator dilution techniques. Workers at the Mayo Clinic used such an indicator dilution technique to measure bile acid secretion and found that it averaged about 12 g/day, rising during meal digestion and falling during the interprandial interval, as well as during overnight fasting (304,305). Bile acid secretion during digestion was about 0.3 µmol/kg-min, a value confirmed in a separate study of gallbladder emptying by G. Van Berge Henegouwen and colleagues, also at the Mayo Clinic (306). Bile acid secretion was also quantified in another intestinal perfusion study by T. Northfield and the senior author, and similar results were obtained (307).

Portal venous blood delivers bile acids to the hepatic sinusoids, and bile acids are efficiently extracted, first pass fractional extraction being 50-90%, depending on the bile acid. C. Einarsson, B. Angelin and their colleagues performed experiments in which they measured bile acid concentrations in portal venous blood and hepatic venous blood in patients at surgery and thus could measure directly hepatic extraction (308). M. Korman and his colleagues found that hepatic extraction of glycocholate was identical during both the fasting and postprandial state (309). This constancy of efficient hepatic extraction means that hepatic extraction is blood-flow limited, and this view was confirmed in a careful study of hepatic bile acid extraction in the dog from the laboratory of R. Hanson (310). First pass clearance for CDCA in healthy volunteers was determined by van Berge Henegouwen and colleagues. They compared the area under the curve (AUC) after intravenous administration, with the AUC after oral administration. A value of 62% was obtained (311). Unconjugated UDCA has a first pass extraction of about 50% (312). In general, dihydroxy conjugates, i.e. the conjugates of DCA and CDCA have a lower first pass extraction than cholyl conjugates, perhaps because the dihydroxy conjugates are bound much more tightly by
albumin. For a given steroid moiety, conjugated bile acids are extracted more efficiently than the unconjugated bile acid.

5. Absorption of bile acids in the biliary ductules; cholehepatic shunting

The discovery that cholangiocytes contain ASBT (238,239) is present in the ileal enterocyte, meant that a fraction of conjugated bile acids secreted into canalicular bile is absorbed during passage through the biliary ductules. The biliary ductules are nurtured by a periductular capillary plexus that empties into the portal region of the hepatic sinusoids. If conjugation is not complete, unconjugated bile acids will enter into canalicular bile. In addition, we found in our laboratory that 24-nor C23 dihydroxy bile acids (norCDCA, norDCA, and norUDCA), which are inefficiently conjugated, are absorbed passively from the biliary ductules (163,164). The absorbed molecules return to the hepatocyte and are once again resecreted into canalicular bile, thus undergoing a cholehepatic circulation. The absorption of the unconjugated bile acid removes a proton from the canalicular bile, generating a bicarbonate anion, the result being no change in bile osmolarity. Lamri et al provided additional evidence for the validity of cholehepatic shunting using immunohistochemical techniques (313). The existence of a “cholehepatic shunt” for unconjugated mono- and dihydroxy bile acids has now been accepted by most hepatologists.

When the unconjugated bile acid returns to the sinusoid, it is taken up by the hepatocyte and resecreted into bile, thereby inducing additional canalicular bile flow, as evidenced by increased mannitol clearance. Mannitol clearance is considered a measure of canalicular secretion (314). The increase in bile flow enriched in bicarbonate is termed “hypercholeresis”. Hypercholeresis was originally described by the laboratory of Serge Erlinger (315). In their experiments in biliary fistula rats, UDCA was infused at a rate exceeding the hepatocyte’s capacity to conjugate it. As a result, unconjugated UDCA was secreted into bile, underwent cholehepatic shunting and induced a bicarbonate rich hypercholeresis. This report from the Erlinger laboratory gave rise to the erroneous belief that UDCA had unique choleretic properties. In cholestatic disease, most patients appear to conjugate their ingested UDCA completely
based on analyses of biliary bile acids (316), but if conjugation were incomplete and unconjugated UDCA were secreted into bile, it would most likely be absorbed in the biliary ductules, thereby undergoing cholehepatic shunting. In further work from our laboratory, we showed that sulindac, but not other NSAID’s, also underwent cholehepatic shunting (317).

6. Determinants of plasma bile acids

The plasma level of bile acids is determined by the instantaneous rates of intestinal absorption and hepatic uptake. Dogs with a chronic biliary fistula have no detectable bile acids in their systemic venous blood, thus confirming that intestinal absorption is the source of plasma bile acids, at least in health (318). The profile of plasma bile acids is not identical to the profile of biliary bile acids. Plasma bile acids are enriched in dihydroxy conjugates because of their lower hepatic extraction.

7. Regulation of the enterohepatic circulation

The enterohepatic circulation of bile acids must balance two conflicting physiological functions. On the one hand, bile acids must serve as the end products of cholesterol metabolism and excretion is necessary for cholesterol balance. For this function, enterohepatic cycling is unimportant. On the other hand, the ileal conservation of bile acids is needed to provide ample bile acids for efficient lipid digestion, and bile acid excretion is unimportant. Given these conflicting aims, it should be highly likely that the enterohepatic circulation of bile acids would be tightly regulated.

The idea of regulation of the enterohepatic circulation is not new. Whipple in his 1922 review (91) noted that there is little increase in total bile acids in dogs ingesting bile acids chronically, and suggested that absorption must be regulated. Sten Eriksson, working in Lund, observed that bile acid synthesis increased with time in the biliary fistula rat (17). Thus the working hypothesis was that bile acid synthesis is determined by the hepatocyte concentration of bile acids. When this falls, bile acid synthesis increases. When bile acids are fed, bile acid synthesis decreases. The senior investigator characterized bile acid metabolism in patients with ileal resection. A marked increase – up to ten fold – was present in
such patients, again attributed to bile acid malabsorption lowering the concentration of bile acids in the hepatocyte (319). A similar increase in bile acid synthesis was induced by cholestyramine administration (320). Studies in the rhesus monkey with an exteriorized enterohepatic circulation of bile acids, also provided convincing evidence that bile acid synthesis was regulated in a negative feedback manner (321). In these studies performed by R.H. Dowling, D. Small, and E. Mack (the surgeon who prepared the monkeys), the return of bile acids to the liver could be systematically decreased, and an increase in synthesis quantified.

However, two observations did not seem to support this useful hypothesis. The first was that bile acid synthesis increased at least acutely when the bile duct is ligated (115). The second, made independently by the group of Reno Vlahcevic in Richmond, VA, (322) and the group of Chijiwa in Japan (323), was that upregulated synthesis in the biliary fistula rat could be returned to normal levels by intestinal, but not intravenous infusion of bile acids. Both types of infusion should lead to an increase in the intracellular concentration of bile acids. The Richmond group proposed that a blood-borne signal from the intestine to the liver might down regulate bile acid synthesis.

The breakthrough came with the identification of FGF15 (in the rat) and FGF19 (in man) as a protein released by the ileal enterocyte in response to FXR induction by bile acids. This observation by Inagaki and Kliewer and colleagues at UT Southwestern (58) was a major advance in our understanding of the regulation of the enterohepatic circulation.

The present view is that the enterohepatic circulation is regulated in two aspects. Input, i.e. bile acid synthesis is regulated at the level of the hepatocyte. Conservation, i.e. ileal transport, is regulated at the level of the ileal enterocyte. Both synthesis as well as ileal transport are downregulated by FGF19 (53). FGF19 enters the hepatocyte via a receptor (FGFR4) and directly or indirectly downregulates bile acid synthesis. Synthesis of FGF19 in the ileal enterocyte is under the control of FXR, the nuclear receptor
activated by bile acids. Nonetheless, in some species, it remains possible that bile acid synthesis is regulated by the bile acid concentration in the hepatocyte rather than by FGF19 (324).

The regulation of ileal conservation is poorly understood. J. Lillienau, working with the senior author, found that bile salt feeding in the guinea pig downregulated ileal transport whereas administrations of a bile acid sequestrant upregulated ileal transport (325). Upregulation of ileal bile acid transport was associated with recruitment of more proximal enterocytes to transport bile acids. Nonetheless, with complete cholestasis or with parenteral alimentation, bile acid transport by the ileal enterocyte decreases (326,327). Thus there must be multiple mechanisms of control. It is likely that the down regulation of the ileal bile acid transporter (ASBT) with bile acid feeding involves FXR agonism and FGF19 release.

C. Perturbations of the enterohepatic circulation in disease

Perturbations of the enterohepatic circulation result from defects in bile acid transport or flow, assuming that bile acid synthesis and conjugation are normal. Bile flow begins at the canaliculus and there can be ductular obstruction, as occurs in cholestatic liver disease such as primary biliary cirrhosis. Physical obstruction can occur with choledocholithiasis or cholangiocarcinoma. Bile flow can be diverted to the outside, as in patients with partial or complete biliary fistula. Bile acids can regurgitate into the stomach and from there into the esophagus. In the small intestine, conjugated bile acids might leak across the epithelium if increased paracellular permeability were present. (To date, such a defect has not been identified). Bacterial overgrowth can be present, leading to deconjugation, and sometimes, to dehydroxylation. Unconjugated bile acids are rapidly absorbed, leading to an intraluminal bile acid deficiency, that in turn can cause lipid malabsorption.

Return of bile acids from the small intestine can be caused by defects in ileal transport, either at the apical or basolateral domains of the ileal enterocyte, or by absence of the ileum. In addition, if rapid intestinal transport propels bile acids past the terminal ileum at a rate exceeding its absorptive capacity, spillover into the colon will occur, and such may induce diarrhea. Portal venous blood flow may be
obstructed or portal-caval shunting may occur. Finally hepatocyte uptake, transport through the hepatocyte, or canalicular secretion may be impaired. It is beyond the scope of this historical review to deal with each of these perturbations on bile acid metabolism.

V. Physicochemical Aspects of Bile Acids

A. Overview

The physical chemistry of biliary lipids and digestive lipids are similar, in that both involve solubilization of polar lipids in the form of mixed micelles. Mixed micelle formation – whether during biliary lipid secretion or triglyceride digestion - have received a great deal of attention from physicians working together with physical chemists, in an attempt to unravel the complexities of these processes.

In 1983, the Kroc Foundation sponsored a small conference (9 physicians, 15 chemists) entitled, “The Physical Chemistry of Bile in Health and Disease”. Many of the topics discussed below are treated in more detail in the published proceedings of that meeting (328). Sadly, it was to be the last Kroc Foundation conference. Ray Kroc, patron of the foundation (and founder of McDonald’s, a global fast food chain) died shortly thereafter, and his widow, Joan Kroc, had other priorities for her generous philanthropy. The foundation was dissolved.

In 1966, the senior author collaborated with Donald Small in the writing of a brief review on the physicochemical properties of bile acids (329). In 1985, Martin Carey authored a systematic review of the physicochemical properties of C24 bile acids (330). His review included tabulation of physical constants, crystal structure as determined by X-ray diffraction, surface chemistry of bile acids, pH-solubility profiles, ionization behavior, and solubilization properties of bile acid solutions. Together, these two reviews summarized existing knowledge of the physicochemical properties of bile acids, as envisioned by medically trained physician-scientists. In the discussion below, only selected aspects of the physicochemical properties are considered.
B. Bile acids in aqueous solution

1. Solubility of the protonated acid

The first detailed studies on the solubility of protonated, unconjugated bile acids were performed by A. Roda and A. Fini at the University of Bologna, Italy (331). They reported solubility values at pH 3 for ten common bile acids. Values varied enormously. That of lithocholic acid was 0.05 µM whereas that of the 7β- epimer of cholic acid (so called “ursocholic” acid) was 1670 µM. Another observation of clinical relevance was that the two bile acids used for gallstone dissolution had quite different aqueous solubilities: that of CDCA was 27 µM whereas that of UDCA was 9 µM. The greater aqueous solubility of CDCA contributes to its excellent gastrointestinal absorption (210); UDCA, with a lower monomeric solubility is absorbed less efficiently and its crystal dissolution is rate limiting in its absorption (332).

The solubility of several protonated glycine conjugated bile acids have also been determined by A. Roda and his colleagues. They are quite low (Aldo Roda, personal communication). Taurine conjugated bile acids, in contrast, even in the acid form, are water soluble.

As yet, we have no solubility values for any of the unconjugated C27 bile acids. The solubility of 5α-cyprinol was found to be 360 µM, thus being similar to that of cholic acid (190). No other solubility data are available for C27 bile alcohols.

2. Ionization

One of the purposes of the Kroc Foundation meeting was to reach a consensus on the pKₐ values of bile acids, as titration studies performed by D. Small (333) and H. Igimi and M.C. Carey (334) gave conflicting results, suggesting that the pKₐ of an unconjugated bile acid was influenced by the number of hydroxyl groups. B. Josephson, working in Stockholm in the 1930’s as a predoctoral student, determined that taurine conjugated bile acids were strong acids, that glycine conjugates had a pKₐ of 4-5, and that
unconjugated bile acids had a pK\textsubscript{a} of about 6 (335). A similar conclusion for unconjugated bile acids was reached by the laboratory of Per Ekwall working in Turku, Finland (336).

At the Kroc meeting, Aldo Roda presented work performed by him and A. Fini which gave accurate pK\textsubscript{a} values for unconjugated bile acids (331). Roda used a technique previously used to determine the pK\textsubscript{a} of long chain fatty acids. One measures the apparent pK\textsubscript{a} in mixtures of increasing proportions of water in methanol and extrapolates to 100% water. The values reported were about 5.1 for the common, natural unconjugated bile acids with no substituents on the side chain. Later, Roda and Fini obtained values of about 4 for glycine conjugates (337). The pK\textsubscript{a} of glycine conjugates is lower than that of unconjugated bile acids, because the electron withdrawal effect of the amide bond enhances the loss of a proton from the carboxyl group (338). The pK\textsubscript{a} of cholestarcosine was determined by our laboratory to be 3.7 (338), about 0.3 pK\textsubscript{a} units lower than that of cholestylglycine.

For unconjugated bile acids, the pK\textsubscript{a} can be reduced by electronegative substituents on the side chain. Roda et al found the pK\textsubscript{a} of 23\textit{R}-hydroxy-CDCA to be 4.8, about 0.2 units lower than that of CDCA (135). It should be possible to lower the pK\textsubscript{a} of unconjugated bile acids considerably by adding one or two fluorine atoms to the C-23 carbon.

The physicochemical significance of conjugation is that it increases solubility markedly at intestinal pH. Dihydroxy bile acids, which constitute 60% of human bile acids, are insoluble at the pH conditions prevailing in the small intestine during digestion (pH 6-7). \textit{In vitro}, the solubility of the bile acid ion increases exponentially with increasing pH. When the concentration of bile acid reaches the critical micellization concentration (CMC), solubility increases markedly, in that a crystalline excess becomes transformed into micelles. There is thus a “critical micellization pH” (CMpH) which defines an approximate concentration above which the bile acid is highly soluble and below which the bile acid is insoluble (339). Lillienau et al, working in our laboratory, found the CMpH of glycocholate to be 4.8 (338). That of CDCA was found by van Berge Henegouwen to be about 6.8 (340) and that of UDCA has
been calculated to be 7.9 (339). The converse of CMpH has been termed “the pH of precipitation” (333) and denotes the pH at which bile acids begin to precipitate from a true solution of the bile acid, the value being dependent on the bile acid concentration. The matter is discussed extensively in the review of Donald Small which gives values for the pH of precipitation of the CDC, DCA, and cholic acid (333).

K.J. Mysels derived the following equation:

$$CMpH = \log [CMC] + pK_a + BApS$$

Where BApS is the negative logarithm of the aqueous solubility of the protonated monomer. The high CMpH for UDCA is thus explained by its low aqueous solubility and a relatively high CMC.

Despite their excellent pH-solubility profile, glycine conjugated bile acids will precipitate in vivo if the pH is sufficiently acidic. This occurs in patients with gastrin-secreting tumors in whom markedly increased gastric acid secretion overwhelms pancreatic bicarbonate secretion and the small intestinal content becomes acidic (341). In the porcupine, glycine conjugated bile acids reflux from the duodenum into the stomach and precipitate, forming gastric bezoars (342).

3. Solubility of Calcium (Ca²⁺) Salts of Bile Acids

Conjugation also greatly increases the resistance to precipitation by divalent cations such as Ca²⁺. Studies from our laboratory (343) in collaboration with K. Mysels (a legendary physical chemist) showed that the calcium salts of taurine conjugates of common natural bile acids are water soluble. Glycine conjugation has a very small effect on the ion product of Ca²⁺ ion x [BA⁻]², and the solubility products of Ca²⁺ salts of glycine conjugated bile acids are quite low. However, the supersaturated solutions of the calcium salts of glycine conjugated bile acids are metastable, resisting formation of the insoluble Ca²⁺ salt for many weeks. The solubilities of Ca²⁺ salts of unconjugated bile acids, like those of glycine conjugated bile acids, are also quite low (344). However, metastability of supersaturated solutions is not observed with unconjugated bile acids.
In the enterohepatic axis, the most dangerous site for possible insoluble calcium salt formation is the gallbladder, where ionized Ca\(^{2+}\) concentrations rise during the concentration of bile by the gallbladder epithelium (345). Stone formation remains rare, as the concentration of ionized Ca\(^{2+}\) is kept rather low in the biliary tract and small intestine. The low concentration of ionized Ca\(^{2+}\) is explained by the high Na\(^+\) concentration, which decreases the activity coefficient of calcium (339).

Formation of Ca\(^{2+}\) bile salt gallstones occurs when “calcium-sensitive” bile acids are present in appreciable proportions in biliary bile acids. In general, this has occurred because of the feeding of unusual bile acids or their precursors (346). Examples are the feeding of cholestanol (the saturated derivative of cholesterol with an A/B trans structure) to the rabbit. The saturated sterol is converted to allocholic acid, that in turn is converted to allodeoxycholic acid by intestinal bacteria. The allodeoxycholic acid is conjugated with glycine and the insoluble glycine conjugate (glycoallodeoxycholate) precipitates from solution forming concretions in the gallbladder (107). Other instances of the formation of insoluble Ca\(^{2+}\) salts are exemplified by the feeding of lithocholic acid to the taurine deficient rat (347,348), as well as the feeding of murideoxycholic acid to the prairie dog (349).

The biological significance of conjugation is that it enhances ionization at the pH present in jejunal content during digestion, preventing passive absorption across the apical membrane of the enterocyte. The bile acid molecule, whether unconjugated or conjugated is too large to traverse the paracellular junctions of the healthy intestine. This “non-absorbability” of conjugated bile acids permits high concentrations of conjugated bile acids to be maintained in the biliary tract and proximal small intestine where the pH ranges from pH 6 to pH 7, and allows them to function as biological solubilizers. Conjugation is also likely to prevent bile acids partitioning into the endoplasmic reticulum during their transit through the hepatocyte.

4. Temperature effects on solubility
Temperature effects are not very important in the biology of bile acids. Solutions of the common primary bile acids are stable at 0° C. Friedrich Krafft, working in Heidelberg, Germany observed that turbid dispersions of soaps became a clear solution over a very narrow temperature range, and his observation has been immortalized in the term “Krafft point” which denotes this temperature of a phase transition. One definition of the Krafft point, is that temperature when the concentration of the bile acid monomer reaches the CMC. With a further increase in temperature, the excess solid phase disappears and the suspension is transformed into a clear, micellar solution. A thorough discussion of the Krafft point can be found in the monograph of Laughlin (350).

The term “critical micellization temperature” (CMT) was proposed, but has not been widely adapted. Sodium cholanoate (no hydroxyl groups) and most mono-hydroxyl bile acids (including lithocholic acid) have high CMT values, well above body temperature.

In contrast, primary dihydroxy- and trihydroxy-bile acid conjugates have very low CMT values. An attempt was made by the senior author to define the CMT values of sodium taurodeoxycholate. This was done by measuring the CMT of sodium taurodeoxycholate:sodium palmitate mixtures with a progressive increase in the mole fraction of sodium taurodeoxycholate. One can then estimate the CMT value of pure sodium taurodeoxycholate by extrapolation. The extrapolated values were well below 0°C (350). Thus, because of their solubility at 0° C, the common natural polyhydroxy C24 bile acids may be considered excellent “cold water detergents”.

The influence of the A/B ring juncture on the CMT of C24 bile acids has not been defined. There is also no information on the CMT values of C27 bile acids as well as those of bile alcohol sulfates with the single exception of 5α-cyprinol sulfate whose CMT was reported to be < 0° C (190).

5. Self-association in solution: Micelle formation

In 1936, G.S. Hartley, working in London, wrote a small lucid monograph entitled, “Aqueous Solutions of Paraffin-Chain Salts: A study in micelle formation”(351). This work summarized the
evidence, mostly based on conductivity measurements, for micelle formation by long chain fatty acid anions as well as alkyl sulfates, and proposed a structure for the spherical micelle. A.S.C. Lawrence, a colleague of Hartley’s once said to the senior author, “Hartley wrote down what everyone else was thinking!” Hartley notes in the final chapter, “The bile salts are obviously amphipathic”. At the end of the Second World War, Hartley returned to academic life and decided to take up the problem of micelle formation in bile acid solutions. However, his postdoctoral student developed mental problems, and the work stopped. Hartley himself left the academic world and became a scientist at Fisons, where he invented novel sprayers for crops as well as developing new pesticides.

The concept of micelle formation had been first advanced by James McBain in 1913, who presented this idea at a meeting of the Royal Society. The chairman listened carefully, and began the discussion with, “Nonsense, McBain.” McBain, who spent the latter part of his scientific career at Stanford, spent much of his subsequent years, characterizing soap solutions, and undertook limited studies with bile acid solutions. McBain also noted in 1941 that detergent solutions can solubilize insoluble dyes. His widow and E. Hutchinson published a book entitled, “Solubilization and Related Phenomena” in 1955, which summarized many of the observations on this topic in the McBain laboratory (352). J. Bashour and L. Bauman, working in New York, developed a method for synthesizing conjugated bile acids, and in 1937 used their synthetically prepared conjugates to show that bile acid solutions solubilize cholesterol in a concentration-dependent manner (353).

In 1942, O. Mellander and E. Stenhagen, working in Sweden, performed conductivity studies with sodium taurocholate and found evidence of aggregate formation (354). Per Ekwall, working in Turku, Finland, summarized the evidence for micelle formation in sodium cholate solutions in 1951 (355). Ekwall also characterized surface properties of bile acids and made many contributions to the physical chemistry of bile acid solutions. (356).
As noted above, a program on the biology and chemistry of bile acids was initiated by Sune Bergström in Lund after the war. Arne Norman synthesized the glycine and taurine conjugates of most of the major bile acids. He then collaborated with Per Ekwall to study the solubilization of 20-methylcholanthrene and thus define the critical micellization concentration (CMC) for dilute solutions of pure, conjugated bile acids (357). Shortly thereafter, and independently, the senior author, working in Lund, Sweden, synthesized and measured the CMC of the major bile acids present in human bile using trans-azobenzene as the solubilizate (358).

In contrast to typical ionic surfactants, the aggregation of bile acid anions is gradual, and there is no well-defined CMC. Thus, the CMC for bile acid solutions is often taken as the midpoint of a small concentration range over which micelles form. K.J. Mysels suggested the term “uncritical micellization concentration”. P. Mukerjee has pointed out that bile acids are rigid molecules in contrast to typical ionic detergents which have a flexible alkyl chain. Therefore, bile acid molecules, just as dye molecules, stack rather than form spherical micelles. The aggregates (from tetramers up to 20mers, depending on bile acid structure and concentration) may be termed multimers rather than micelles. The size of the aggregates was estimated by light scattering, especially by D. Small (329,333) and J. Krahtohvil (359). However, as noted below, aggregates of bile acid molecules alone, i.e. without an accompanying lipid, do not occur in nature, except when phospholipid extrusion into bile is defective.

The most recent measurement of CMC values for a large number of unconjugated and conjugated bile acids was made by Aldo Roda in our laboratory, in collaboration with K. J. Mysels (360). This study was successful because of two major advances. The first was the synthesis of a large number of bile acid epimers by Takashi Iida and Frederick Chang (361), and in a few cases, by our laboratory. The second was the availability of a maximum bubble pressure device designed, constructed, and characterized by K.J. Mysels. A key aspect of this technique for measuring surface tension in relation to concentration, is that a new bubble surface is constantly being formed whose composition consists of the dominant constituent of the solution. In contrast, with surface tension methods using a tensiometer, a more surface
active impurity may concentrate at the surface of the solution to be analyzed and lead to erroneous values for the CMC. Roda validated his bubble pressure values by showing that identical values were obtained with the technique of dye solubilization using Orange OT (1-O-tolyl-azo-2-naphthol) as the water-insoluble dye. Values are reported in the absence and presence of 0.15 M Na⁺ concentration.

The senior author, when working in Sweden in the 1960’s observed that the aggregation of monoglycerides and bile acids to form mixed micelles was cooperative, so that solubilization occurred at a concentration well below the CMC of bile acids in the absence of such amphiphilic lipids (358). W. Duane used dialysis to measure the intermicellar concentration of bile acids (monomers plus simple micelles) in both model bile samples and human bile samples. He obtained values for the critical micellization concentration as well as for the intermicellar concentration of bile acids which were quite similar to the critical micellization concentration in bile acid-monoglyceride systems (261).

C. Solubilization of polar lipids by bile acid micelles.

1. Physical chemistry of biliary lipids

Bile acids are present at micellar concentrations only in bile and small intestinal content. In man, the concentrations in plasma (2-10 µM) and in colonic content (400 µM) are well below the CMC. In health, bile acids are always associated with a polar lipid. In bile, this is predominantly phosphatidylcholine (PC). In small intestinal content, during digestion, the polar lipids are monoacyl glycerol and fatty acids (partly ionized). The only circumstance in which bile acid micelles are present without an accompanying polar lipid is in the Mdr2 knockout mouse and in patients with a defective Mdr3 protein. Such mice (and patients) lack this ATP-activated PC extruder in the canaliculus of the hepatocyte and are believed to have essentially no phospholipid in ductular bile. The micellar bile acids, if sufficiently hydrophobic, attack the membranes of the cholangiocytes, leading to ductal inflammation in mice (362) and to a variety of biliary problems in patients (363). Nonetheless, the bile of many vertebrate species has an extremely low ratio of phospholipids to bile acids (181) without any sign of biliary ductule injury. In such species, it
may be that phospholipid, after secretion into canalicular bile, is absorbed in the biliary ductules with the result that it is not present in gallbladder bile in appreciable proportions, or that the ductules are highly resistant to attack by bile acids or both.

The recognition that lecithin was present in bile occurred early in the 20th century. G. Whipple in his 1922 review (91) tabulated the reported analyses of biliary lecithin. In 1952, Polonovski and Bourillon, working in Paris, confirmed that the dominant phospholipid in bile was lecithin, later to be named PC (364). G. Phillips separated biliary phospholipids chromatographically and confirmed that the dominant phospholipid in bile was PC (365). Another article in this series gives the history of phospholipids in more detail.

Franz Ingelfinger, one of the dominant figures in American Gastroenterology in the mid 1900’s was unhappy with the level of ignorance of the pathogenesis of the common disease of gallstones. He arranged for Donald Small, one of his brightest fellows, to spend two years in Paris working under the direction of a scholarly colloid chemist, Digren Dervichian. During those two years, the behavior of model systems simulating bile was described clearly using triangular coordinates. The ternary systems – water-lecithin-bile acid, water-cholesterol-bile acid, water-cholesterol-lecithin – were carefully defined using sodium cholate, egg lecithin, and cholesterol. The behavior of the three lipids and water could then be described by a tetrahedron. A slice across the tetrahedron at 10% total lipid concentration gave a triangle with three coordinates – cholesterol, lecithin (PC) and bile acid. A line defined the limits of the micellar zone. Accordingly, points lying above the line would be supersaturated in cholesterol either in liquid crystalline or crystalline form. This classic work is summarized in reference 333 and illustrated in Figure 9. The stage was set for applying this pure exercise in model systems to the pathogenesis of cholesterol gallstone formation (see later).

That phospholipids play a key role in the solubilization of cholesterol in bile was perhaps first recognized in the mid 1950’s by B. Isaksson, working in Sweden (366). Work on the solubilization of
lecithin by bile acid solutions was carefully studied by S. Yoshimuta (367), a member of the Department of Surgery at Kyushu University in Japan whose director, E. Miyake, had a longstanding interest in gallstone formation. This work was published in the 1960’s about the same time that D. Small was studying model systems simulating bile in Paris. Another physician with interest in this problem was Kerrison Juniper who attributed cholesterol solubilization in bile to a macromolecular complex (368). A review by D. Small and the senior author, published in 1967, advanced the idea that the physical chemistry of micelle formation in bile and micelle formation in small intestinal content during triglyceride digestion were similar physicochemical processes (329).

Thus by 1967, both bile and small intestinal content were considered micellar solutions, the only difference being the nature of the accompanying lipids. Simple bile acid micelles are poor solubilizers for cholesterol or non-polar dietary lipids. The mixed micelle provides a hydrocarbon interior in which poorly soluble lipids can dissolve. In bile, this insoluble lipid is cholesterol; in small intestinal content, insoluble lipids of physiological importance are the fat soluble vitamins. The remaining vexing problem was the molecular arrangement of the mixed micelles present in bile and small intestinal content.

2. Molecular arrangement of mixed micelles

To solve the problem of the structural arrangement of the mixed micelles in bile, D. Small traveled to England to the laboratory of Dennis Chapman (F.R.S.), a physical chemist with a deep interest in membrane structure. They examined solutions by NMR and found that the angular carbons of the bile salt molecule were in a hydrophobic environment. On the basis of the NMR spectra, they proposed the “mixed disc” model in which a bilayer of PC molecules with a drum shaped perimeter was coated on its surface with bile acid molecules (369). Later, dimers of bile acids interdigitating between the PC bilayer were added (330). The model was widely accepted.

At the Kroc meeting in 1984, G. Lindblom of the University of Lund proposed that bile acids could also interact with phospholipid bilayers by lying flat in the bilayer with the hydrophobic face of the bile
acid molecule pushing the charged heads of PC apart (370). W. Nichols and J. Ozarowski proposed that bile acids lie flat on the surface of worm-shaped micelles (371) in contrast to the mixed disc model. Rex Hjelm, a biophysicist working at the Los Alamos National Laboratory, examined bile acid-PC-water systems using small angle neutron scattering. Hjelm concluded that as the PC/bile salt ratio increased, the change in scattering indicated that PC-bile acid micelles increase in size by the elongation of constant diameter rods (372), providing additional support for the “radial shell model” structure advanced by Nichols and Ozarowski. Hjelm also concluded that the disk-shaped micelle originally proposed by Small et al. could not be correct. Nonetheless, the conclusions of Hjelm are not widely known in the medical community, because much of his work was published in chemical journals that are not included in the PubMed data base. A depiction of the original model for the mixed micelle proposed by Small and Carey is compared with that of the model inferred from the small angle neutron studies of Hjelm and his colleagues in Figure 10.

In 2002, S. Marrink and A. Mark, working in the Netherlands used the technique of molecular dynamic simulations, in which molecules aggregate in water to give the lowest free energy (373). They concluded that, “phospholipids are packed radially with their headgroups at the surface and the hydrophobic tales [sic] point…. toward the micellar center. The bile salts act as wedges between the phospholipid headgroups”. They concluded that the structure of the micelles generated in their system resemble the radial shell model proposed by Nichols and Hjelm. Additional work by Long et al (374) using small angle X-ray scattering and quasi-elastic light scattering were also in agreement with the model proposed by Nichols.

Another approach to study the interaction of bile acids and phospholipids was taken by Heuman et al (375). They prepared single layer vesicles and characterized the binding of bile acids to them. They also determined at what phospholipid/bile acid ratios, the vesicles were transformed into micelles.
For teaching purposes, it would be nice to have a video with realistic molecular models showing the transformation of lipid bilayers into mixed micelles by bile acid anions.

D. Physical chemistry of bile and cholesterol gallstone disease

1. Supersaturated bile in bile from gallstone patients

D. Small returned from France to Boston in 1965, and began to apply his studies on model systems simulating bile to human biliary lipid compositions. Bile from radiolucent gallstone patients and healthy control subjects was collected. The proportions of bile acid, PC, and cholesterol were plotted using triangular coordinates in a study in which W. Admirand did much of the leg work. A line denoting equilibrium cholesterol saturation had been obtained by D. Small based on his studies in Paris: samples above the line were supersaturated with cholesterol whereas those below the line were unsaturated in cholesterol. When the data were plotted, bile from patients with gallstones was supersaturated with or without cholesterol microcrystals, whereas bile from most non-gallstone patients was unsaturated (376).

[The paper also re-introduced triangular (or ternary) coordinates to the medical audience. Triangular coordinates can be used when there are three variables adding up to a constant sum, i.e. there are two degrees of freedom, permitting the constructions of graphs in two dimensions. They were originally proposed in 1873 by the great physical chemist J. Willard Gibbs, of Yale University and two decades later popularized by a Dutch chemist, H.W.B. Roozeboom (377). Probably, their first usage in medical journals was in a paper by E.H. Ahrens (378) published in 1957 to describe the proportions of fats, carbohydrates, and protein in diets].

Cholesterol gallstone disease could now be considered a disease of defective bile secretion by the liver, rather than being the result of a diseased gallbladder, as had often been proposed in the past. The paper also indicated that biliary lipid composition should be expressed as a ratio of solute (cholesterol) to solvent (bile acid-PC micelles). The story of the rise and fall of medical dissolution of cholesterol gallstones is discussed later.
2. Vesicle formation and gallstone disease

In the original phase diagram of D. Small, the physical state of cholesterol in supersaturated systems depended on the ratio of PC to bile acid. At low ratios, cholesterol was crystalline. At higher ratios, the excess was liquid crystalline, i.e. in lamellae of PC and cholesterol. Norman Mazer and Martin Carey used quasi-elastic light scattering to characterize model systems simulating bile. At higher PC to bile acid ratios, they found evidence for “microprecipitates” although they also suggested that these microprecipitates were vesicles enriched in cholesterol compared to mixed micelles (379,380). About the same time Giora Somjen and Tuvia Gilat, working in Israel, used gel permeation chromatography to separate mixed micelles from vesicles (381). They equilibrated their columns with 10 mM cholate prior to chromatography and thus the vesicular phase remained unchanged during the chromatographic separation. They also noted that if the columns were preequilibrated with buffer alone, a bile sample would become entirely vesicular, as the bile salt molecules moved into the aqueous spaces in the gel permeation column, thereby increasing the phospholipid/bile acid ratio. Thomas Holzbach, working in Cleveland, also found vesicles in human bile samples and showed that nucleation of cholesterol crystals from the cholesterol-rich vesicles occurred rapidly (382). William Higuchi, working in Salt Lake City, showed using dialysis techniques that model bile systems contained four constituents: monomers, simple bile acid micelles, mixed micelles (containing bile acids, PC, and cholesterol) and vesicles also containing the three biliary lipids. The vesicles were relatively enriched in cholesterol (383).

Because of the complexity of the physicochemical aspects of bile and because of the widespread interest in gallstone pathogenesis and treatment, in 1989, the NIH sponsored a workshop entitled, “Biliary Cholesterol Transport and Precipitation”. The proceedings (including much animated discussion) were published as a supplement to Hepatology the following year (with financial support from the Ciba-Geigy Corporation) (384). The meeting included a wide range of scientists from pure membrane biophysicists to gastroenterologists and surgeons, and the discussion often indicated problems in communication. A summary of the meeting was authored by Steven Strasberg, a scholarly surgeon, and his associate Robert
Harvey. Their summary (385) presented a reasonably coherent view of bile supersaturation. Cholesterol and phospholipids are secreted in vesicular form. Bile acids, initially present as simple bile acid micelles, solubilize phospholipid and cholesterol. However, the transfer process, presumably occurring by collision, solubilizes phospholipids preferentially, leaving behind vesicles containing an excess of cholesterol (from which cholesterol crystallizes). Thus, there is a progressive change in the physical state of bile as it moves from the canaliculus into the gallbladder.

At this meeting, J. Donovan, then working in the Carey laboratory, tabulated the existing clinical data (386). The proportion of cholesterol in vesicles was higher in hepatic bile than in gallbladder bile in all studies, indicating that the transformation of vesicles to mixed micelles continued during gallbladder storage of bile. C.Schriever and D.Juengst showed, using bile samples from gallstone patients, that the ratio of cholesterol to phospholipid in vesicles purified by gel filtration was 2.0 (387), a ratio that favors formation of cholesterol crystals.

Since this meeting, especially because of the declining interest in medical dissolution of gallstones, work on the physical chemistry of bile has waned. W. Higuchi continued his work on model systems, emphasizing the “activity” of monomeric cholesterol (388). Cholesterol saturation continued to be calculated from the original triangular figure of D. Small, even though the limit to the micellar zone was moved downward, based on in vitro studies of model systems from the laboratories of Henrik Dam in Copenhagen (389), and that of R.T. Holzbach in Cleveland (390). [Henrik Dam, a Dane, had a long interest in cholesterol. He received the Nobel Prize in 1943 for his discovery of vitamin K. In the 1960’s, he turned his enormous energy to the study of cholesterol gallstones and did both animal (391) and clinical studies (392) of note. Sadly most of his publications are to be found in an obscure journal (Z. fur Ernahrungswissenschaft) and are seldom cited]. Simple methods for calculating cholesterol saturation based on the revised line of saturation were developed. M. Carey published “critical tables” that give a value for cholesterol saturation for any molar ratio of phospholipids to cholesterol, as well as varying total lipid concentration (393). The line of saturation is also influenced by bile acid type -- being lower when
bile acids are enriched in conjugates of UDCA. M. Carey has also published critical tables for this special circumstance (394).

E. Physical chemistry of fat digestion

1. Studies on model systems simulating triglyceride digestion

The physical chemistry of model systems simulating fat digestion has received much less attention. The senior author, when working in Sweden, defined the behavior of each of the relevant constituents in dilute bile acid solutions. Fatty acids (oleic) were readily available, but monoacyl glycerides (monoglycerides) were not. Fortunately, they were generously supplied by Fred Mattson, a chemist working at the Miami Valley laboratories of Proctor and Gamble in Ohio. [Mattson, a gifted biochemist, discovered the positional specificity of pancreatic lipase (395) and then showed that triglyceride was absorbed from the intestine in the chemical form of monoacylglycerol and fatty acid (396).] The solubility of fatty acids in bile acid micelles was pH dependent, as protonated fatty acid has a rather low solubility in bile acid micelles, whereas fully ionized fatty acid (soap) is infinitely soluble (350). Simple titration experiments (397) indicated the pKₐ of fatty acids in a taurine conjugated bile acid micelle was about 6.9, indicating that fatty acids at proximal small intestinal pH (where most fat absorption occurs) are about half ionized. Mono-olein, whether the 1- or 2-isomer, was highly soluble, with the ratio of solubilized monoolein to micellar bile acid approaching 1.8 (358). In contrast, di- and triacyl glycerides had little solubility in bile acid micelles.

In 1988, M. Svärd working in the group of Björn Lindman in Lund, investigated the taurocholate-monolein-water system by a great variety of techniques, and constructed the complete phase diagram (398). In her paper, she also noted minor, but consistent differences between this system and the bile acid-PC-water system elucidated by the pioneering work of D. Small two decades before. What both diagrams share in common is a large micellar zone, as shown for the taurocholate-monolein-water system.
illustrated in Figure 11. The paper of Svärd et al also proposes that bile acid molecules lie flat on the surface of the bilayers.

2. Presence of a micellar phase in vivo during triglyceride digestion

The in vitro findings indicated that fatty acids and monoglycerides were highly soluble in bile acid micelles, but that di- and triglycerides were not. These observations led to the prediction that during fat digestion, glycerides and fatty acids would be in two phases. There would be an oil phase containing di- and tri-acyl glycerides and a micellar phase consisting of bile acid micelles containing partly ionized fatty acid and monoglyceride. The oil phase might also contain protonated fatty acids if these were generated by pancreatic lipolysis at a rate exceeding that which could be solubilized in mixed micelles. Later, healthy volunteers were intubated, fed a mixed meal, and intestinal content was collected. It was then ultracentrifuged at body temperature, and both the oil phase and micellar phase analyzed. The results confirmed the predictions derived from the in vitro experiments (399).

Some years later, a more detailed analysis of this system was performed by O. Hernell and J. Staggers working in the M. Carey laboratory. They studied both model systems (400), as well as small intestinal content obtained by intubation in healthy volunteers who had ingested a triglyceride rich meal (401). They confirmed the older studies of the senior author, but noted in addition, that vesicles were present. They proposed that vesicles form at the surface of the oil droplets, and that these are fairly rapidly converted to mixed micelles when they encounter unsaturated bile acid micelles. In 1959, A.S.C. Lawrence, a colloid chemist at Sheffield University, proposed that detergents adsorb to “dirt” and form liquid crystalline myelin figures (multilayered vesicles). His analysis (402) is probably an apt description of the formation of vesicles composed of fatty acid and monoacylglycerol at the surface of the triglyceride droplet.

A closing chapter on this topic was authored by R. Hjelm, in which model mixtures simulating fat digestion were prepared in our laboratory and examined by small angle neutron scattering. Hjelm
concluded that mixed micelles containing a bile acid and either monoglyceride (monoacylglycerol) or monoglycerides plus fatty acids had essentially the same structure as the mixed micelles containing bile acids and PC that he had studied previously (403,404). In both, the bile acid molecules lie flat on the surface of the spherical or worm shaped micelles.

3. Necessity of bile acids for efficient absorption of triglycerides

Numerous studies dating back to the 19th century had shown that triglyceride absorption efficiency was impaired in animals with a biliary fistula and that feeding of bile acids would correct fat malabsorption (405). Nonetheless, there were efforts to show that bile acids were unnecessary for fatty acid absorption (406). The view of the senior author is that bile acids are essential for the efficient absorption of those fatty acids that have very low aqueous solubilities, i.e. saturated fatty acids of C₁₆ length or longer (407). Fatty acids with appreciable aqueous solubility (unsaturated fatty acids and saturated acids having a chain length < C₁₆) do not require bile acids for efficient absorption. This view that bile acids facilitate the absorption of saturated fatty acids (≥ C₁₆ in length) is based on two observations. First, fecal fatty acid analyses indicate that exogenous bile acids cause a decrease in the proportion of long chain saturated fatty acids in short bowel syndrome patients (408). Second, administration of cholestyramine to rats causes selective malabsorption of long chain saturated fatty acids (409). It is well established that bile acids are required for the absorption of fat soluble vitamins.

The lack of requirement of bile acids for efficient absorption of fatty acids with appreciable aqueous solubility provides the rationale for using medium chain triglycerides for nutritional support in patients with a deficiency of bile acids in the small intestine. The liberated medium chain fatty acids are quite water soluble and are rapidly absorbed in the complete absence of bile acids.

VI. Dissolution of cholesterol gallstones by oral bile acids and topical solvents

A. Gallstone dissolution by oral bile acids.
As noted in the introduction, the famous triangle of D. Small (and W. Admirand) indicated how biliary lipid composition could be plotted. (The term “biliary lipids” to include bile acids, PC (lecithin), and cholesterol was probably introduced by D. Small in the 1960’s. It did not include bilirubin diglucuronide (which gives bile its color), as this pigment derived from heme is a very minor constituent of bile. Biliary lipids differ from plasma lipids in that biliary lipids include bile acids, and do not include glycerides and fatty acids).

In the 1950’s, Charles Johnston, a surgeon at Wayne State University, was interested in determining what he termed the “cholesterol holding capacity” of bile, that is, how much additional cholesterol could be dissolved when added to bile samples. He observed that T-tube bile from gallstone patients was “saturated”, whereas bile samples obtained from healthy subjects undergoing surgery was unsaturated. With a young Japanese surgeon, Fumio Nakayama, he fed cholic acid to gallstone patients, but found that it did not increase the “cholesterol holding capacity”. In the discussion of his paper, he noted that he hoped to try CDCA, but of course, at that time, there was no CDCA to be had (410). [Fumio Nakayama continued to study gallstone pathogenesis using chemical techniques and had a long and distinguished career as a Professor of Surgery in Kyushu, Japan. After he retired he summarized his lifelong interest in gallstone formation and treatment in a monograph (411)].

After J. Thistle and L. Schoenfield made the seminal discovery that CDCA but not cholic acid changed bile from supersaturated to unsaturated, it was logical to perform a clinical trial of CDCA for gallstone dissolution. The effort was spearheaded by J. Thistle, as L. Schoenfield took a new position in Los Angeles in 1970. In 1971, the first observations were made that gallstones were decreasing in size, probably the first witnesses being J. Thistle (and his radiological colleague) and the senior author. In 1972, successful gallstone dissolution was reported in May at the annual meeting of the American Gastroenterological Association and later that year in Europe at the 2nd International Bile Acid Meeting sponsored by the Falk Foundation of Freiburg, Germany. The early Mayo experience with gallstone dissolution was published in the New England Journal of Medicine in 1972 and became a “Citation
Classic” (37). Later that year, gallstone dissolution by CDCA was reported by the group of R.H. Dowling, working at Guy’s Hospital in London (38). In 1973, Dr. Falk Pharma decided to market CDCA for gallstone dissolution.

In February 1972, a small meeting was held at the NIH to discuss a national multicenter placebo-controlled trial of CDCA. It was felt that the absence of patent protection would preclude development by pharmaceutical companies. A “Request for Proposals” was issued in July 1972. L. Schoenfield submitted a proposal which included committee structure for the proposed National Cooperative Gallstone Study (NCGS). On the committees were most every American scientist or physician interested in gallstone disease!

The study was delayed, because of the FDA concern that CDCA might cause liver damage. The FDA demanded that liver biopsies be performed on the first 100 patients, including those on placebo. Despite the ethical problems inherent in subjecting placebo group patients to liver biopsy, the study was performed without any adverse events.

The concern of the FDA was based in part on two reports from the Huntington Laboratories in England of severe toxicity induced by CDCA feeding in the rhesus monkey (412). As a result of these reports, European studies were stopped in 1973. Report of the primate toxicity provided a justification for liver biopsies in the patients when therapy was stopped (413). An international registry of biopsies was set up, and a careful review of the biopsy histology showed no important histological changes. Trials in Europe recommenced.

The answer to this remarkable species difference in CDCA toxicity came from work by the senior investigator and his colleagues who showed that in man, lithocholic acid, the major bacterial metabolite of CDCA, is sulfated and rapidly excreted from the body (414,415), whereas in the rhesus monkey, lithocholic acid is extremely poorly sulfated, accumulates in the circulating bile acids, and induces liver damage (416,417). Independent studies by the lab of E. Mosbach reported that CDCA also was toxic in
the rabbit, again because of lithocholic acid accumulation (418). Similar findings were observed when rhesus monkeys (419) and baboons (420) were given CDCA.

The NCGS began in 1976 and its results were published in 1981 (421). The study was a double blind study with three groups -- placebo, 350 mg CDCA/day and 750 mg CDCA/day. Dissolution occurred in the placebo group (11%), was greater (24%) in the low dose group and highest (41%) in the high dose group. During the study it was recognized that the high dose was probably still too low a dosage (422), but the study could not be changed. In the high dose group, 8% of patients had to discontinue the study because of an increase in serum aminotransferase levels. Toxicity was shown to be attributable to CDCA and not to lithocholic acid (423). While the NCGS was proceeding, controlled studies were being performed in Europe which showed that CDCA had unequivocal efficacy and minimal toxicity (424,425).

An analysis of the kinetics of gallstone dissolution by William Higuchi indicated that the reduction in diameter induced by the presence of unsaturated bile occurred independently of size (426). Thus, the bigger the stone, the longer the time needed for complete dissolution. This prediction was later confirmed by Senior et al in a thoughtful analysis of x-ray changes of gallstone size during medical dissolution (427).

Just as the NCGS was beginning, word came from Japan (a careful study led by Isao Makino and his colleagues) that ingestion of ursodeoxycholic acid, the 7β- epimer of CDCA, would also induce gallstone dissolution (40). Earlier, there were several anecdotal cases in which subjects had taken large doses of UDCA and induced their own gallstone dissolution, and this was done without knowledge of the American studies with CDCA. Thus, medical dissolution of gallstones must be considered an American and Japanese discovery.

In the United States, the NCGS drafted a protocol that would permit a study similar to the original NCGS, but would compare UDCA with CDCA; however, the study was not funded by NIH. CDCA was approved for marketing in the United States by the FDA in 1983, and UDCA was approved a few years
later. UDCA, like CDCA, was also toxic in the rhesus monkey (428) and rabbit (429) because of lithocholic acid accumulation in the circulating bile acids.

Despite the safety and efficacy of CDCA and UDCA, there were some physicians who were unimpressed by medical therapy as compared to surgical therapy. There were several problems with medical therapy. First, it took a long time – months to years -- to obtain complete dissolution, especially for larger stones. Second, studies by Wolpers suggested that with the first attack of biliary colic, the resulting inflammation would cause a surface layer of non-cholesterol constituents on the gallstone, slowing or precluding dissolution (430). Thus, there was no window of opportunity: when stones became symptomatic, they had already become resistant to medical dissolution. Moreover, it took many months of therapy before one could be sure that the stone was resistant. Finally, after stones dissolved, there was recurrence (431). It was gallstone recurrence that led surgeons a half century before to switch from cholecystotomy (removal of the stone but leaving the gallbladder in place) to cholecystectomy (432). A thoughtful review by Gracie and Ransohoff concluded that it was better to “wait and cut” than to “screen and treat” because of the benign natural history of most gallstones (433). An editorial by K. Isselbacher was entitled “Disillusion with Dissolution” (434).

**B. Acceleration of dissolution by extracorporeal shockwave lithotripsy**

Attempts were made to speed up dissolution. One approach was to fragment the stones by using extracorporeal shockwave lithotripsy, a technique that had been developed for the fragmentation of kidney stones. For gallstones, the stone fragments would be smaller than the original stone(s), have cholesterol surfaces exposed by the shattering of the stones, and thus, medical dissolution would occur more rapidly. However, there was a major difference between kidney stones and gallbladder stones. When kidney stones were fractured by this technique, the fragments passed spontaneously down the urinary tract and were expelled during urination. With gallbladder stones, the fragments sunk to the bottom of the gallbladder, so that dissolution by desaturating bile acids was needed. Another limitation of
extracorporeal shock wave lithotripsy was that only certain types of stones (one or two in number) could be treated. Thorough clinical studies of the efficacy and safety of this technique for gallstone dissolution were conducted by the Munich group led by Gustav Paumgartner (435). Leslie Schoenfield led one of several efforts in the United States (436).

C. Topical dissolution with organic solvents

Another approach to non-surgical treatment of cholesterol gallstones was contact (topical) dissolution with organic solvents. The Mayo group, led by J. Thistle, introduced the idea of placing a percutaneous transhepatic catheter into the gallbladder and then lavaging the stone with methyl tert-butyl ether (MTBE) a solvent that had become available as a gasoline additive (437) and was a superb cholesterol solvent. When the solvent was pumped into and then aspirated from the gallbladder, stones would dissolve in a matter of a few days. A spark-free pump was designed and patented by the Mayo Clinic. European trials commenced in a few centers (436,437).

In our laboratory at UCSD, we had not planned to become involved in this approach. A patient with symptomatic gallstones who was a poor operative risk presented with symptomatic gallstones, and we reluctantly carried out successful dissolution by manual instillation and aspiration of MTBE. A “smart” pump was designed by S. Zakko so that solvent could be pumped into and out of the gallbladder at a high rate. The pump was programmed to continuously sense intraluminal pressure and aspirate gallbladder contents whenever the gallbladder started to contract, thus avoiding spillage of the solvent into the common bile duct (438). A new solvent – ethyl propionate – was tested and found to dissolve gallstones as nearly as rapidly as MTBE (439). Ethyl propionate did not induce nausea in contrast to MTBE because of rapid removal from plasma (440). Patients were treated successfully (441).

However, there were multiple problems with contact dissolution. The solvents used were commercial samples and they were not manufactured according to FDA standards. MTBE was flammable and considered toxic (442). Preclinical studies had not been done. The procedure required some hours for
complete dissolution, and as described, it required professional personnel to perform the infusion and aspiration. In studies performed in pigs, the gallbladder epithelium was observed to be destroyed by either solvent, although it recovered rapidly (443). There was a report from France of MTBE escaping into the blood stream and causing severe renal damage (444). In the United States, two deaths occurred unrelated to usage of the solvent, but nonetheless in association with solvent usage. A company was formed to bring the pump and solvent into clinical practice. However, venture capital was never obtained, and the approach slowly passed into oblivion. A review of topical dissolution by our group was published (445).

D. Concurrent advances in gallstone management

Two other developments were occurring during this time that doomed widespread usage of medical dissolution. The first was the development of laparoscopic cholecystectomy, a major surgical advance. Laparoscopic surgery was an import from France, used first in gynecological procedures. Using this technique, a gallbladder could be removed quickly, without a linear scar, and the patient could leave the hospital in a few days. The gallbladder with its gallstones had been removed, and the patient was cured. In experienced hands, the technique was as safe as the traditional open cholecystectomy which had been performed widely for the past half century. The composition of the stone(s) was unimportant, whereas with oral bile acid therapy only gallstones that had cholesterol as the major constituent could be successfully treated...

The second advance was ultrasound imaging of the gallbladder and its contents. All types of stones were detected, and there was no radiation exposure. In contrast, oral cholecystography required the ingestion of iodine-rich contrast agents and subsequent X-ray to detect the presence of gallstones.

Thus in twenty years, medical dissolution had been discovered, prospered, and then slowly disappeared. In the long history of biliary disease, medical dissolution was a “bubble”. Once again symptomatic cholelithiasis became a surgical disease. Nonetheless, medical dissolution was “proof of
principle”, and medical therapy with UDCA can be used to prevent cholesterol gallstone formation. Gallstone prophylaxis might be useful in some clinical situations, for example, in patients with recurrent gallstone pancreatitis who did not wish to have surgery.

The availability of UDCA led to the accidental discovery of its utility in the treatment of primary biliary cirrhosis and cholestasis of pregnancy, as noted in the introduction.

**VII. Epilogue: omissions and apologies**

Space limitations means that historical aspects of many facets of biology and chemistry have been omitted. We have not discussed the history of inborn errors of bile acid synthesis and conjugation. Our understanding of these genetic defects is quite recent, and an excellent review (446) by Clayton is available. In the United States, Ken Setchell, a talented mass spectroscopist, has described several new inborn errors of bile acid biosynthesis, as well as a defect in bile acid conjugation. Identification of the exact enzyme defect requires mass spectrometry to detect the presence of bile acid precursors and gene sequencing to identify the defect (156). The rewards of discovery are great for the patient, as bile acid replacement may be life-saving. Some of the early history is given in J. Sjövall’s retrospective review (10).

A second area not discussed is the utility of bile acids and bile alcohols to disclose evolutionary relationships, an area pioneered by G.A.D. Haslewood (21). Often bile acids agree with the perceived evolutionary relationships, but sometimes they do not. It would be of interest to explore these nonconcordancies in detail.

A third area not discussed is the role of bile acids in colonic secretion (447,448) and absorption (449), as well as recent work suggesting that bile acid activation of TGR5 is essential for normal defecation, at least in the mouse (70). Details of the biochemical mechanism(s) for bile acid-induced intestinal secretion are being elucidated by several laboratories, among these being that of Stephen Keely, in Dublin, (448) and of Meena Rao in Chicago (450). Bile acid malabsorption causes diarrhea, and this can
be treated by sequestrant administration (451,452). A new syndrome of defective FGF19 release causing an inappropriate increase in synthesis of bile acids has been identified by Julian Walter and his colleagues (453), and is considered to be present in many patients with irritable bowel syndrome of the diarrhea type (454), as well as in patients with so-called primary bile acid malabsorption. Plasma levels of “C4” (7α-hydroxy-cholest-4-ene-3-one), an intermediate in C24 bile acid biosynthesis, have been shown to correlate highly with the rate of bile acid synthesis (455,456). The necessity of measuring C4 (as a marker of increased bile acid synthesis) and FGF15/19 levels (to detect impaired FGF15/19 secretion from the ileal enterocyte) in every patient with unexplained diarrhea is slowly being recognized by gastroenterologists (457). The group of Michael Camilleri at Mayo is exploring the relationship of fecal bile acid profile to constipation as well as diarrheal conditions (458). Our laboratory identified some children with functional constipation whose fecal bile acids were predominantly the 3 sulfate of CDCA (459). The absence of C-12 hydroxylation in these children could indicate a deficiency in cyp8B1, the enzyme that hydroxylates C4 at the C-12 position.

A fourth area not discussed is the cellular toxicity of bile acids. Mechanisms of toxicity are multiple, involving activation of the death receptor (460), as well as induction of reactive oxygen species (461). UDCA is cytoprotective and mechanisms responsible for this effect are being studied (462). A new aspect of bile acid toxicity is the proinflammatory action of retained bile acids (463).

One last area not discussed is the possible influence of bile acids on the microbiome. The presence of secondary bile acids in bile has long been a sign that metabolic activity of the microbiome influences bile acid metabolism. Indeed, characterization of bile acid metabolism in the germ free animal (464) by Gustafsson and his colleagues in Lund aided in the initial distinction of primary and secondary bile acids by the laboratory of Sune Bergström (8). There is still much is to be learned about the effect of individual bile acids on the microbiome (465).
The last two generations have had powerful tools to elucidate bile acid metabolism. One of these was the discovery of $^{14}$C – a discovery that almost occurred by chance (466). The availability of $^{14}$C in chemical form permitted $^{14}$C labeled bile acids to be prepared and their metabolism to be defined. A second essential tool has been chromatography – adsorption, reversed phase (HPLC), and gas/liquid. A final tool is mass spectrometry which when coupled with gas chromatography or HPLC provides both qualitative and quantitative information on new molecules (9). For the assignment of structure to newly discovered bile acids, proton and $^{13}$C nuclear magnetic resonance measurements are mandatory. Of course, in today’s research, generation of specific knockouts, characterized by multisystem phenotyping, combined with PCR, have enormous power in unraveling the recently discovered signaling properties of bile acids.

At present (2014), at least two types of drugs based on natural bile acids are being tested for their utility in the treatment of liver disease. The first, obeticholic acid, the $6\alpha$-ethyl derivative of CDCA, is a potent FXR agonist. Its efficacy in the treatment of primary biliary cirrhosis (467) and nonalcoholic steatohepatitis (468) is being tested in clinical trials. NorUDCA, the C$_{23}$ homologue of UDCA, is also in clinical trials for the treatment of cholestatic liver disease. In the bile duct ligated mouse, its feeding causes less damage than UDCA feeding (469).

Inhibitors of ASBT, the sodium dependent bile acid transporter present in the apical domain of the ileal enterocyte, are once again being developed. They were originally aimed at providing adjuvant therapy for hypercholesterolemic patients who responded inadequately to statins (470). Now the target is cholestatic liver disease or its symptom of pruritus as well as type 2 diabetes (471) and constipation (472). Colesevelam is an improved bile acid sequestrant that should have a similar pharmacodynamic activity as that of the ASBT inhibitors; however, it is only approved for type II diabetes, and its use to treat bile acid diarrhea is off label. Bile acid derivatives that activate solely TGR5 or both TGR5 and FXR have been synthesized (55) and may eventually enter clinical trials.
In our opinion, the remarkable progress in our understanding of bile acid biology and chemistry is a tribute to the hundreds of scientists who have worked with such diligence and talent throughout the world. We have mentioned only a few in our brief history, and we have surely omitted some key advances. There are many workers whose contributions have not been recognized because of space limitations. As we noted in the introduction, this is a personal history, and other writers are likely to stress different aspects of the historical record. We have probably overemphasized the contributions of our own laboratory because we know them best. For those readers who would like more historical detail on some of the topics discussed above, we refer them to a series of beautifully written historical reviews by Adrian Reuben published in the journal *Hepatology* between 2002 and 2006.

Let us hope that future research in the chemistry and biology of bile acids will be as exciting as that of the last 80 years!
Acknowledgments: We are grateful to many of our colleagues who helped us in this review. In particular, we thank Takashi Iida, Bruno Stieger, and David Mangelsdorf, as well as the anonymous reviewers. We also acknowledge a grant-in-aid from Intercept Pharma, Inc. In the past, our laboratory has been supported by the National Institutes of Health as well as a grant-in-aid from the Falk Foundation, e.V., Freiburg, Germany.
REFERENCES


Mataki, M. Pruzanski, R. Pellicciari, J. Auwerx and K. Schoonjans. 2009. TGR5-mediated bile acid

2007. The G-protein coupled bile salt receptor TGR5 is expressed in liver sinusoidal endothelial cells.
*Hepatology* **45**:695-704.

70. Alemi, F., D.P. Poole, J. Chiu, K. Schoonjans, F. Cattaruzza, J.R. Grider, N.W. Bunnett, and C.U.
Corvera. 2013. The Receptor TGR5 mediates the prokinetic actions of intestinal bile acids and is

receptor. Gpbar1 (TGR5), negatively regulates hepatic inflammatory response through antagonizing


73. Sobotka, H. Physiological chemistry of the bile. 1937. Williams and Wilkins, Baltimore, MD.


Amsterdam, The Netherlands.


77. The Bile Acids. Volume 1, Chemistry, 1971; Volume 2, Physiology and metabolism, 1973; Volume 3,
NY.


Kritchevsky, and P.P. Nair, eds. The Bile Acids: Chemistry, Physiology, Metabolism, vol. 4. Plenum


**72**:9298-307.


Pharmacie.* **149**: 185-201.

Fuchsbichler, J. Sjövall, and M. Trauner. 2006. FXR(-/-) mice adapt to biliary obstruction by enhanced

canalicular transport and toxicity of dietary ursodeoxycholic acid in the abcb11-/- mouse: transport and


196. Stahl, E. Dünnschichtchromatographie. 1962. Springer Verlag, Berlin, Germany


homogeneous fluorescence enzymatic assay, and isotope dilution gas chromatography-mass spectrometry. 


methyl tert-butyl ether or ethyl propionate in animals: A comparison of two solvents for contact


stimulates Cl(-) secretion via cAMP signaling and increases cystic fibrosis transmembrane conductance


**3**:349-57.


LEGENDS TO FIGURES


Figure 2. Numbering system of bile acids and bile alcohols, and frontal view of CDCA. The numbering system was developed in the 1930’s when the steroid structure was finally established.

Figure 3. Molecular structure of cholesterol (above) and CDCA (below) showing the changes in the cholesterol molecule when a C24 bile acid is formed. The changes are (1) reduction of the double bond to give a 5β (A/B cis) A/B ring juncture (2) a modified β- oxidation of the side chain to remove three carbon atoms and convert the terminal carbon to a carboxyl group; and 3) epimerization of the 3β- hydroxy group to a 3α- hydroxy group. For a discussion of the enzymes involved, see reference 155.

Figure 4. Depictions of the two epimers of decalin, a saturated C10 hydrocarbon that is used to illustrate the two A/B ring juncture epimers of bile acids. When the two bridgehead hydrogen atoms are across from each other, the juncture is called trans; when they are on the same side of the ring juncture, the juncture is called cis. When bile acid structure is shown in frontal view (Figure 2), the only indication of the stereochemistry of the A/B ring juncture is the orientation of the hydrogen atom, i.e. whether alpha or beta. However, in bile acids, the substituent at the upper bridgehead carbon atom is the C-19 methyl group. Often bile acid structure is shown without any indication as to the orientation of the hydrogen atom at C-5. When this is done, it is understood that the molecule is A/B cis, as 5β- (A/B cis) bile acids.
are much more common in nature than 5α (A/B \textit{trans}) bile acids. A/B \textit{trans} bile acids are termed “allo” bile acids.

\textbf{Figure 5.} Silhouette of a C_{24} conjugated bile acid molecule emphasizing its possession of a hydrophobic side (beta face) and a hydrophilic side (alpha face). Common sites of hydroxylation in addition to the default structure (3α, 7α) are at C-6, C-12, and C-16; all of these are on the hydrophilic side of the molecule. Modified from Roda et al (360).

\textbf{Figure 6.} Time course in healthy subjects of levels in peripheral venous plasma of immunoreactive conjugates of cholic acid (cholyl conjugates) and those of CDCA (here termed “chenyl” conjugates) in response to meals; meals are indicated by the thick, vertical arrows. Levels of CDCA conjugates rise sooner than those of cholyl conjugates indicating more proximal absorption. Levels of CDCA conjugates are also higher than those of cholyl conjugates despite similar proportions in bile because of lower fractional extraction by the liver. From reference 216.

\textbf{Figure 7.} Schematic depiction of the work of L. Lack and I. Weiner on conjugated bile acid (here, taurocholate) absorption by everted sacs of rat intestine, showing preferential and active absorption by the distal ileum. At time zero, bile acid concentrations are identical on both sides of the intestine. Modified from reference 229.

\textbf{Figure 8.} Depiction of cholesterol and bile acid metabolism as well as the enterohepatic circulation by Bergström, Danielsson, and Samuelsson in 1959 (8). In the accompanying text, they note that the mass of cholesterol entering the small intestine in bile is greater than that of dietary cholesterol. The figure shows neither the spillover of absorbed bile acids into the systemic circulation nor selective absorption of conjugated bile acids by the terminal ileum. Values for biliary bile acid secretion and fecal bile acid excretion are about two fold higher than those obtained in more recent studies.

\textbf{Figure 9.} Depiction of the use of triangular coordinates to plot biliary lipid composition and thereby to define the saturation of cholesterol in bile. The triangle is derived from a plane parallel to the base.
indicating a constant water composition (10% solids, 90% water) in the tetrahedron consisting of the four biliary components (water, lecithin, cholesterol, and bile acids). The micellar zone is shown in black. Bile having a lipid composition falling within in the micellar zone is unsaturated. If the composition is above the upper limit of the micellar zone, the sample is supersaturated in cholesterol and at risk for cholesterol gallstone formation (376). If biliary lipids are known, the per cent saturation can also be obtained from tables published by Carey (393,394).

**Figure 10.** Molecular arrangement of the drum shaped mixed micelle micelle originally proposed by Small (369) and modified by Carey (330), as compared to the radial shell model proposed by Nichols and Ozarowski (371). The radial shell model was confirmed by Hjelm et al using small angle neutron scattering (372,403,404) and also by molecular dynamic studies of Marrink and Mark (373).

**Figure 11.** Phase diagram of the water-monoolein (monooleylglycerol)-taurocholate system, showing the large micellar area (cross-hatched), thus depicting the excellent solubilizing potency of bile acids for polar lipids. Liquid crystal states are indicated by vertical and horizontal cross hatching. Modified from Svärd et al. (398).
Table. Common and uncommon sites of additional nuclear hydroxylation (in addition to the default structure of the 3α,7α-dihydroxy compound) in primary C$_{24}$ bile acids, C$_{27}$ bile acids, and C$_{27}$ bile alcohols

<table>
<thead>
<tr>
<th>Class</th>
<th>Common</th>
<th>Uncommon+</th>
<th>Site</th>
<th>Natural occurrence</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>C$_{24}$ bile acids</td>
<td>6α,6β,12α,16α</td>
<td></td>
<td>1α</td>
<td>Australian opossum</td>
<td>474</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1β</td>
<td>Pigeons, neonates</td>
<td>475, 476</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2β</td>
<td>Neonates</td>
<td>477</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4β</td>
<td>Chinese pheasant</td>
<td>473</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9α</td>
<td>Bear</td>
<td>478</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11α</td>
<td>Sunfish</td>
<td>472</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15α</td>
<td>Swan, goose</td>
<td>478, 479</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19*</td>
<td>Neonates</td>
<td>480</td>
</tr>
<tr>
<td>C$_{27}$ bile acids</td>
<td>12α</td>
<td></td>
<td>1α</td>
<td>Tinamou</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16α</td>
<td>Ancient birds</td>
<td>473</td>
</tr>
<tr>
<td>C$_{27}$ bile alcohols</td>
<td>12α</td>
<td></td>
<td>2α</td>
<td>Arapaima</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6α and 6β</td>
<td>Elephants, hyrax, manatee</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11α</td>
<td>Sunfish</td>
<td>473</td>
</tr>
</tbody>
</table>

+ Hydroxylation occurs at C-5 with 24-nor (C$_{23}$) UDCA in hamsters (481).

*The bile acid described is a tetrahydroxy bile acid (3α,7α,12α,19-tetrahydroxy-).
Figure 4

Trans-Decalin

Cis-Decalin
Figure 5
Figure 6
Figure 8
Figure 9
Figure 10
Figure 11